cultural science

Analysing Geo-linguistic Dynamics of the World Wide Web: The Use of Cartograms and Network Analysis to Understand Linguistic Development in Wikipedia [1]

Han-teng Liao
Institute for the Study of Diplomacy, Georgetown University, Washington, USA

Thomas Petzold
ARC Centre of Excellence for Creative Industries and Innovation, Queensland University of Technology, Brisbane, Australia

ABSTRACT

This article discusses the usefulness of geo-linguistic analysis for Internet studies by presenting two techniques to frame and visualize the linguistic development of the World Wide Web, in particular the geo-linguistic development amongst different language versions of Wikipedia. An emergent research agenda has been set to explore the multilingual aspects of the Internet using, for example, a global perspective on Wikipedia research. And yet, there is a lack of theoretical and methodological tools for understanding the distribution and diffusion of linguistic materials online. The idea of geo-linguistic factors is introduced in this article to address these shortcomings and to respond to the study of a wide range of issues such as linguistic pluralism on the Internet or, more generally, the diffusion of innovation. Cartograms and network analysis are presented as two techniques that showcase the potential uses of geo-linguistic analysis. These two techniques of measurement and visualization indicate certain geographic and linguistic affiliations among languages. It is argued that although certain more developed language versions such as English and German may have central positions in connecting all languages, there exists another pattern that can best be explained by geo-linguistic factors. Finally, the limitations and implications of such findings and techniques are discussed, not only for research on Wikipedia but for Internet studies in general.

Introduction: Geo-linguistic Factors and Their Use

Media research about globalization and localization raises the questions about how media institutions and audiences are reorganized and realigned along or across boundaries. Meanwhile, the software that underpins computers, digital networks, and the Web has undergone a process of internationalization and localization that facilitates the adoption of related technologies for users of various languages and from different regions.[2] Leading global search engine Google, for example, provides different interfaces and tools for more than 170 local domains and 120 languages and variations such as Canadian English, U.K. English, and U.S. English.[3] Similarly, Wikipedia, the global user-created encyclopaedia, has over 270 language versions, including improved flexible interfaces for variations of Chinese (Mainland simplified, Singapore/Malaysia simplified, Hong Kong traditional, and Taiwan orthodox) within the Chinese version of Wikipedia (Liao, 2009).[4] Thus, it could be expected that this process of internationalization and localization (which support the everyday interaction and usage in the digital environment) has also reorganized and realigned institutions and users in different ways. However, there is little research in this area that brings the seemingly "technical" issues into the supposedly "macro" media research about globalization and localization. Yet, their everyday impact on activities in the digital environment cannot be overlooked. Such neglect may be explained by the early monolingual (English only) development of the Internet. There is, however, little excuse for researchers now to ignore the fact that languages (along with regional factors) have increasingly become a significant aspect of the Internet. Therefore, while the factors of languages and regions are indeed important for media research in general and for Internet research in particular, how researchers can approach these with appropriate working concepts and cutting-edge techniques remains an open issue.

This article suggests that geo-linguistics (or geography of language), a small but emergent field that exists between socio-linguistics and human geography, may provide some concepts and techniques for Internet and media researchers to explore the political, cultural, and social implications of internationalization and localization in the digital environment. Geo-linguistics is concerned with the distribution of languages over time and space. Thus, geo-linguistic analysis can assist researchers and policy makers by critically investigating the "what, where, when, who, and why" (Cartwright, 2006) questions about languages at "international (macro), national (meso), and urban (micro) levels" (van der Merwe, 1993, p. 23). Indeed, some media scholars have taken similar paths in the past using the concept of "geolinguistic regions" (Albizu, 2007; Sinclair, 1996) to explore the role of languages and regions in areas such as international and national TV programming. This article argues that geo-linguistic analysis can also inform our understanding of the linguistic development of the Internet, and associated temporal and spatial changes. Some new kind of geo-linguistic analysis designed for online linguistic development should be expected to generate useful insights for future research and policy making, and may provide tangible research findings to better understand the relationship amongst globalization, localization and the Internet. This article initiates research endeavours in this emerging field by exploring the findings of a preliminary study on the geographic and linguistic affiliations amongst Wikipedia language versions. The results are expected to be relevant for researchers to reconsider traditional issues such as identity politics, language politics, and media marketing (that addresses markets across various languages and regions) in the digital environment.

Geo-linguistic factors, be they implicit or explicit, have sometimes facilitated and sometimes hindered certain technology adoption. Users can both be empowered and conditioned by the geo-linguistic configuration offered in their respective digital environments. For example, Japanese-speaking travellers may have difficulties locating Internet cafés abroad that provide a Japanese-ready environment, that is, where they can read and type Japanese language correctly. Furthermore, users can find their search results more rapidly and efficiently if they can manipulate the geo-linguistic conditions of the digital environment. It might be a better strategy, for example, to find British government-related information by searching the British version of Google (google.co.uk) instead of using any other language version. Moreover, search activities depend on the users' capacity to type keywords in a certain language, which are submitted to certain search engines that may provide different services by considering the users' geographic and linguistic affiliations (e.g., Australian or Mainland China users who want to circumvent certain government-initiated restrictions pertaining to accessing particular websites). Therefore, Internet researchers must take geographic and linguistic factors seriously as they play an increasingly important role in software development as well as in everyday interaction in the digital environment.

Wikipedia as a Critical Observation Site of the World Wide Web

Wikipedia can be regarded as one of the most interesting sites of observation for issues pertaining to geo-linguistic analysis online. Firstly, it claims to be the "free encyclopedia" that anyone can write and edit. With its increasing number of language versions, Wikipedia is argued to be the most comprehensive website containing different linguistic materials in a single site. Therefore, it should be rich enough in linguistic materials for researchers to understand the geo-linguistic development of the global Wikipedia project as well as the larger World Wide Web. Secondly, because each linguistic version is governed and run by its editors, its language policy debate and development can be different from those traditional language-planning settings (Liao, 2009). Thirdly, because of the "inter-language links" that aggregate and connect all language versions of the same (sometimes only similar) entry, Wikipedia contains rich information about how and when certain languages are connected and linked, providing some indication about the development, diversity, and diffusion of the linguistic materials it aggregates and maintains.

Geo-linguistic analysis is useful to answer questions on the diffusion and diversification of languages within Wikipedia and, more generally, the World Wide Web. How Wikipedia handles linguistic diversity and how its exponential growth spreads geographically are but two pertinent questions that can be asked within a geo-linguistic framework. Wikipedia's geographic development enables us, for example, to understand its (lack of) popularity in different regions around the world that may be explained by (a combination of) economic, political, social, cultural, and (in particular) linguistic reasons.

The geographic development of Wikipedia has reached a point of saturation in some regions, most notably in parts of Europe and the United States. These are the areas where Wikipedia cannot expect further staggering growth in numbers but rather, in regional and linguistic diversity, stability and quality. This will affect some but not necessarily all of

the original seven language versions (that reached more than 100 articles in the year of Wikipedia's foundation in 2001): English, German, Spanish, Polish, Portuguese, Dutch, and Swedish. Wikipedia's diversity depends on its diffusion into different regions and areas around the world. On its official statistics website (http://stats.wikimedia.org), Wikipedia currently lists more than 270 languages in total, and categorizes them into six of the world's regions plus one category of constructed language.[5] At the core of its geographic growth and diffusion, we can begin to discern Wikipedia's linguistic policy from the broader linguistic development.

A critical analysis of geographic and linguistic affiliations of any site of investigation on the Internet (and elsewhere) requires a thorough understanding on how the units of analysis come into existence. Thus, in order to understand geo-linguistic activity on Wikipedia we need to understand the linguistic policy process behind it, in other words how Wikipedia languages get approval or become rejected. A multiple-step process determines the addition or rejection of a new language version to Wikipedia. The application procedure is extensively documented and can be monitored at all times.[6] The specific requirements Wikipedia has for a new language proposal to be approved are:

1. A new language edition must not already exist on any project of Wikimedia (of which Wikipedia is arguably the most popular one).

2. The language must have a valid ISO-639 1–3 code.[7]

3. The language must be sufficiently unique that it could not coexist on a more general wiki.[8]

4. A sufficient number of living native speakers form a viable community and audience.[9]

Wikipedia's linguistic policy has adapted a traceable language adoption procedure so as to involve the community and concerned individuals. This has allowed Wikipedia to grow and diversify and essentially to become the most linguistically diverse project in digital culture.[10] And yet, we may still ask how geo-linguistically distributed and diverse Wikipedia really is.

Geo-linguistic Analysis for Diversity and Diffusion Measurement

Wikipedia is not fully available to many people around the world. At the moment, it provides a platform for five percent of the world's 6,000+ languages. It comes as no surprise then that Wikimedia's top two emerging strategic priorities focus on expanding within large as well as midsized and under-connected populations (Wikimedia, 2009). While language versions of large populations (such as China and India) promise exponential growth in numbers, reach remains an issue (e.g., Wikipedia's Chinese competitors Baidu and Hudong). Even when diffusion (of Wikipedia) may be increased by its reach to Chinese and Indian language populations, the level of linguistic diversity does not necessarily increase accordingly.

For one thing, the rise of major Chinese and Indian languages (Mandarin Chinese and Hindu) does not guarantee the online development of other dialects and minority languages that exist in China and India. Paolillo (2007) asserts that the "Internet could shift over to Chinese as the dominant language and not become any more linguistically diverse in the process" (p. 424). He points out that whilst Mandarin Chinese accounts for fifteen percent of the global population, simply adding Chinese languages could actually decrease linguistic diversity (ibid). It needs to be stressed that such empirical approaches (quantitative and qualitative) are rarely employed in debates on linguistic diversity and, even if they are, the process of information gathering is further complicated by inconsistent and outdated statistics (cf. Gerrand 2007).

Linguistic diversity, however, must not be confined to the mere nominal listing or enumeration of languages. Such common practice of measuring linguistic diversity (counting the number of languages within a given sphere), should be considered only as a standing point. Further understanding is required to see its actual activities, where a participatory digital culture can provide some indication. A participatory understanding can most notably explain how access to and participation on the Web is relevant for a specific language and its speakers (and how it is not), and why languages with fewer speakers are potentially and actually disadvantaged against major languages (and why they are not). The participatory and networking potential of digital culture is generally expected to provide new data, practice and ideas to advance the ideal of linguistic diversity. Still, it has been argued elsewhere that what constitutes sustainable linguistic plurality in digital culture is not the mere participation of the world's 6,000+ languages, but their mutual interaction—36 million language pairs is the ultimate extrapolation from current possibilities (Petzold, 2010). Researchers have to discern the current reality from the future potential. The authors thus believe that "participatory understandings" as in the debates on digital divides and digital literacy must be complemented by understanding some realities about digitally-enabled language interaction. Geo-linguistic analysis provides techniques that enable us to critically investigate such activities. There is much to be gained from measuring the interaction and diffusion among languages online.

Two Techniques of Geo-linguistic Analysis

There exist many techniques which combine geographic and linguistic factors that can help researchers to explore, connect, and understand geo-linguistic research. It is outside the scope of this article to provide a comprehensive review of these techniques, but two categories of such possibilities can be identified by the way geographic and linguistic factors are combined. One category is to make maps about languages by showing and analyzing the linguistic data on maps. Another category is to conduct network analysis about the relationships between languages by showing and analyzing their interconnections. Using Wikipedia's data, this article demonstrates a few cartograms and a network map, for each category respectively, to show the potential usefulness and pitfalls for further detailed analysis in the future.[11]

Choropleth maps and cartograms: distribution, diversity, and diffusion

One conventional way to show linguistic data on a map is to make a choropleth map. In a choropleth map, areas are labeled, colored or patterned in correspondence to the

statistical variable of interest. A straightforward way to show Wikipedia's linguistic development across the world is to draw a choropleth map to show certain indicators of development. For example, if the numbers of articles, editors, and so on can be used as a proxy for the development of certain language versions of Wikipedia, then a choropleth map can be useful to demonstrate the geographic distribution and possible affiliation of different Wikipedia projects. In Figure 1, a choropleth map of European national languages shows the number of Wikipedia articles of their associated national languages in graded colors, with English (en) as darkest color and Irish (ga) as brightest.

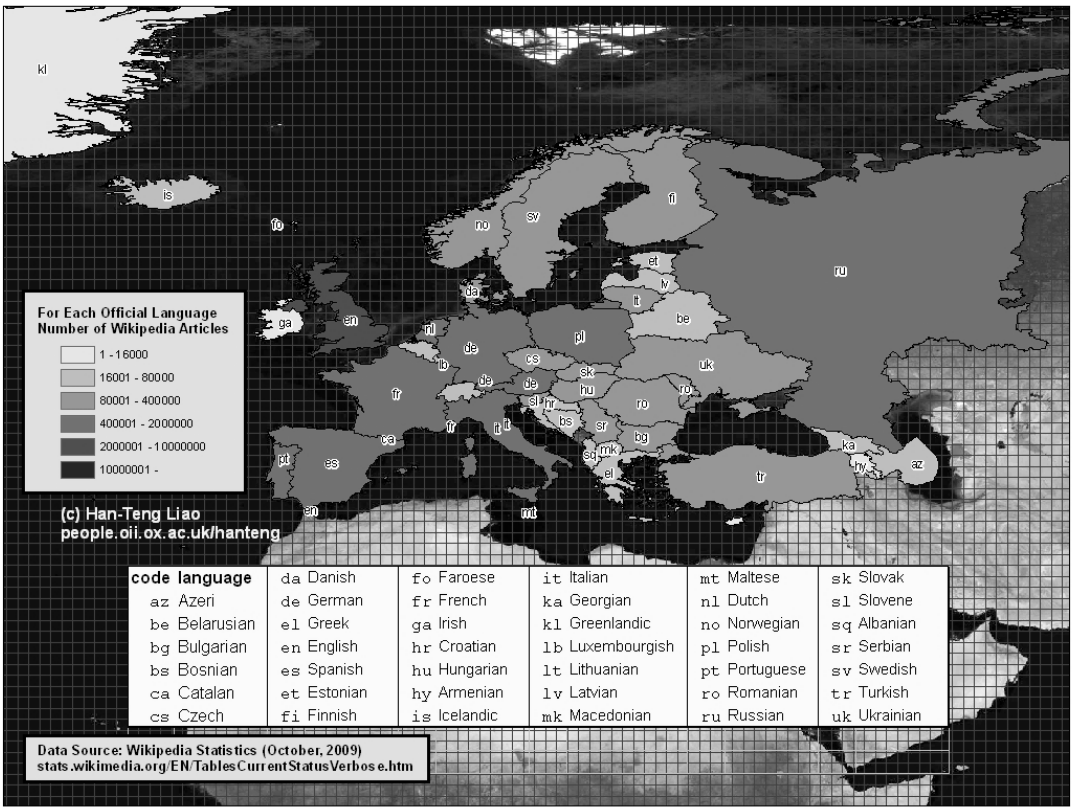

Figure 1: A choropleth map showing the number of Wikipedia articles for each national language in Europe

However, it might be controversial to map a certain language onto one or several regions since the idea of one region with one language may not be applicable. Indeed, the belief that one region, especially one nation, has one language is contested, and is related to issues such as linguistic nationalism (mobilizing linguistic resources for nationalist agendas) and multilingualism (more than just one major linguistic capacity in a community). For example, not all nation states have only one official language (e.g., Belarus has Belarusian and Russian; Ireland has Irish and English; and states like Belgium, Switzerland, and Cyprus have official languages that are official languages of other states). Also, some nation states share the same official language, such as German in Austria, Germany, Liechtenstein and Switzerland. Due to the limitation of mapping along the boundaries of nation states, the creation of the following map adhered to certain "mapping rules" which results in each nation state only having one official language at present on the map. After all, it is as if we are assigning existing Wikipedia language

versions to each nation state. Therefore, it should be noted that by creating maps for this paper we had to make a few decisions in order to connect official languages to nation states, a procedure which is obviously open to other arrangements and criticism. Firstly, it is straightforward to assign the region to the official language that no other nation state uses. Secondly, for those states which do have more than one official language, and if some of the official languages are already assigned (e.g., Russian for Belarus and English for Ireland), they are deliberately removed from that region. The results are that all European official language versions of Wikipedia are assigned, with Belgium, Switzerland, and Cyprus left out.

By considering the geographic factors in the choropleth map in Figure 1, it is relatively intuitive to see that English is the dominant language in the Wikipedia project while other western European languages seem to have substantial numbers of articles as well. The next tier is filled with Nordic languages and some official languages of Eastern Europe. In turn, the next layer has languages such as Estonian (et), Latvian (lv), Belarusian (be), Croatian (hr), Bosnian (bs), Albanian (sq), Macedonian (mk), Icelandic (is), Azeri (az), and Georgian (ka). Language versions that stand out on this map in terms of geographic affinity are possibly Russian (ru) and Polish (pl), with more Wikipedia articles than the Nordic countries where Internet and information and communication technologies (ICT) development is expected to be more extensive. With this simple map, researchers can generate some ideas on how certain geographic and linguistic factors may influence the distribution or adoption of Wikipedia projects in Europe, and also consider the level of linguistic diversity inside the Wikipedia project for Europe.

Using the same dataset for a cartogram (Figure 2) shows a somewhat "distorted" map to readers. Indeed, the basic idea of cartograms is to substitute geographic properties such as area and distance with the variables of interest, so that the map appears distorted. Still, cartograms should not be regarded as deformed designs of maps, but rather as an alternative way to present a certain aspect of reality. As a graphical method it shows "the pattern of distribution of a single element" (Raisz 1938, p. 256) as opposed to a topographic map, which is made up of a combination of elements. For example, area cartograms are increasingly popular mapping techniques of showing the size of regions proportional to other different kinds of datasets, for example, GDP or population.[12] In other words, areas are adjusted to reflect the corresponding variables. As shown in Figure 2, due to the nature of cartograms, it becomes immediately obvious that English is dominant in Wikipedia's world of European official languages, with its large area representing a strong presence. Still, German, Polish, Dutch, and various Latin languages have strong showings considering their geographical size. In contrast, Russian has a relatively small size of Wikipedia articles in comparison with its geographical territory.

By making a choropleth map and a cartogram with Wikipedia's dataset, researchers can begin to generate some preliminary observations and hypotheses, largely about the distribution and diversity of geographic and linguistic components of the global Wikipedia project in Europe. If similar map-making processes are repeated with Wikipedia's dataset for each year, the temporal elements can be included and examined. This article argues that such a technique is useful for studying the diffusion of innovation (within Wikipedia in this case) in order to consider whether geographic and/or linguistic affinity may have helped or hindered certain Wikipedia language versions to evolve. The

geographic and/or linguistic affinity used in this paper refers to the connection and relationship that can be explained or shown by geographic and/or linguistic factors.

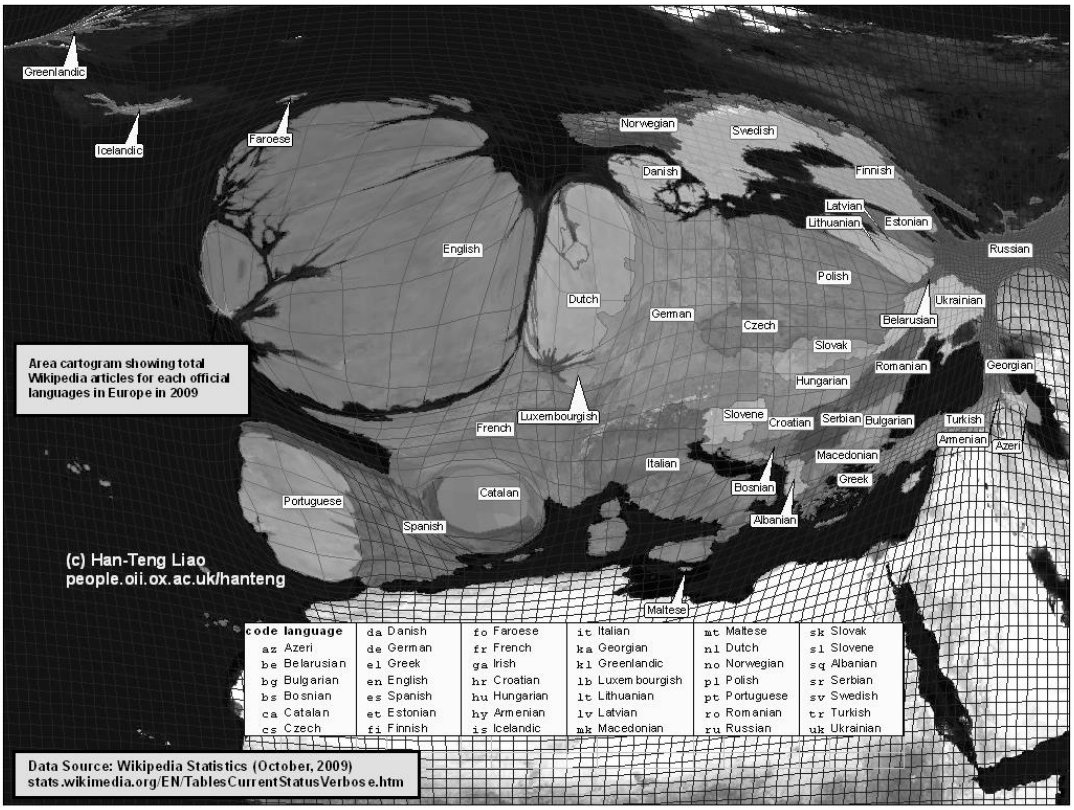

Figure 2: A cartogram showing the number of Wikipedia articles for each national language in Europe

It is further suggested that such mapping of geo-linguistic factors of certain media development throughout a period of time can advance the concept and related discussion of "geolinguistic regions" (Albizu, 2007; Sinclair, 1996) for media research. It is particularly crucial to do so for Internet research because, unlike the areas of TV programming and film where national and regional boundaries can largely be assumed, websites such as Wikipedia and Google may not have clear segmentations when it comes to regions and languages. It remains an open question whether websites such as Wikipedia and Google have created certain "geo-linguistic regions" as in the research on global television. Still, with the techniques described above, it is feasible to collect certain datasets (web pages, users, activities, etc.) for a longitudinal study of the development along geo-linguistic lines. In this way, researchers may gain insights about not only "what" kinds of geo-linguistic regions have emerged but also "how" they have evolved.

This article particularly assumes that the use of cartograms can help researchers to understand the process of diffusion, an interdisciplinary concern categorized under the umbrella term of "diffusion research" (Bruce & Yearley, 2006). It is worth mentioning that the "diffusion research" in sociology, media studies, human geography, ethnography, etc. is borrowed from the concept of diffusion in physics. It is then understandable that

computer programs, which make cartograms manageable, have been involved with the concepts of "distribution," "flows," and "diffusion." For example, Tobler (2004) describes the objective of cartograms within a physical analogy: "One may imagine that a thin sheet of rubber is covered with an uneven distribution of inked dots representing a distribution of interest. The objective is to stretch the rubber as much as necessary until the dots are evenly distributed on the sheet" (p. 67). Moreover, "deliberate distortion occurs in order to make room for the symbols on an illustration depicting flow" (p. 58). The algorithm that created cartograms shown in this paper is designed by physicists who use a "diffusion-based" method to tackle computation problems (Gastner & Newman, 2004). It therefore seems appropriate to use area cartograms for diffusion research. For example, the cartograms of Figure 3 and Figure 4 indicate that the spread of Internet use in East Asia starts mainly from regions such as Japan, Korea, and the three Chinese-speaking regions of Hong Kong, Taiwan, and Singapore. Finally, it can be speculated that the growth of Internet users in mainland China may have been caused by its geographic affinity to all these regions plus a linguistic affinity with Hong Kong, Taiwan, and Singapore, especially with the large Internet population in the southern and coastal provinces of mainland China.

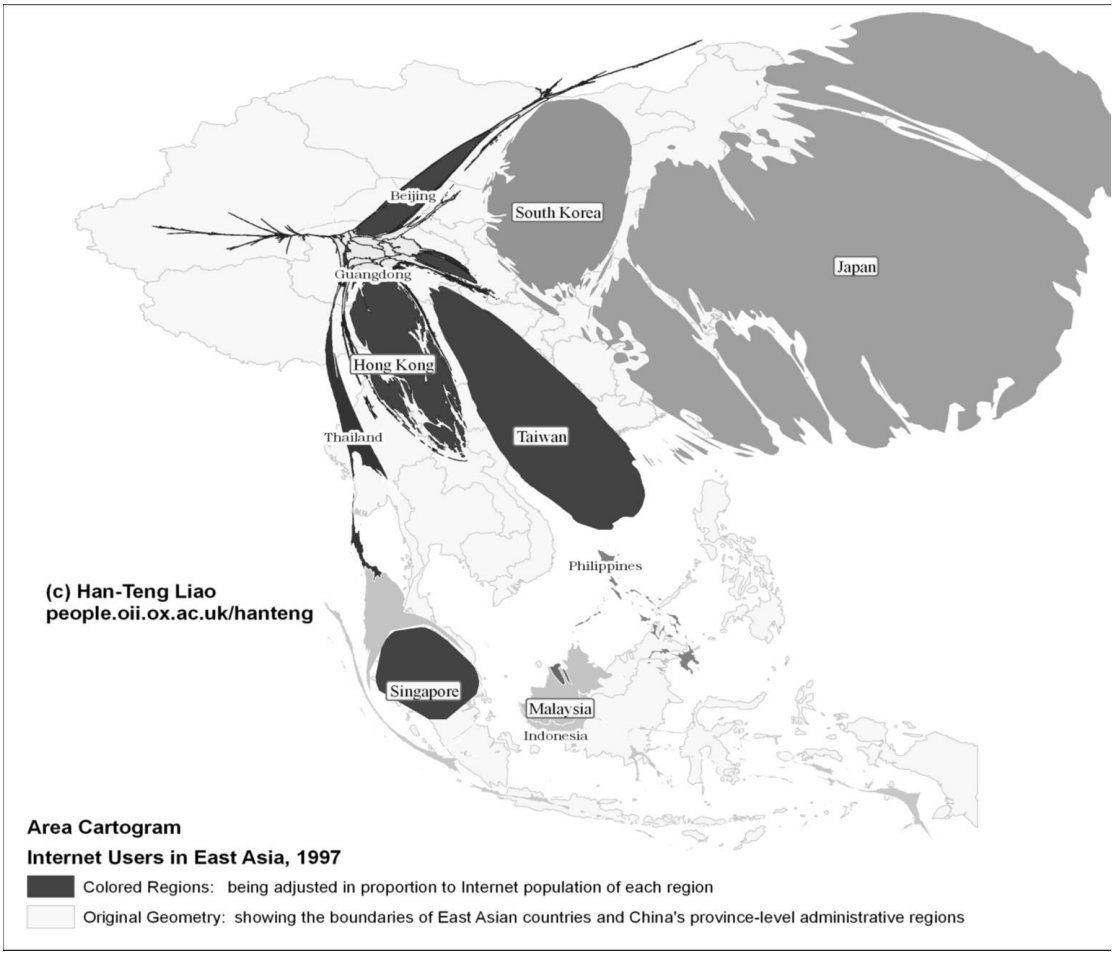

Figure 3: A cartogram of Internet users in 1997 East Asia

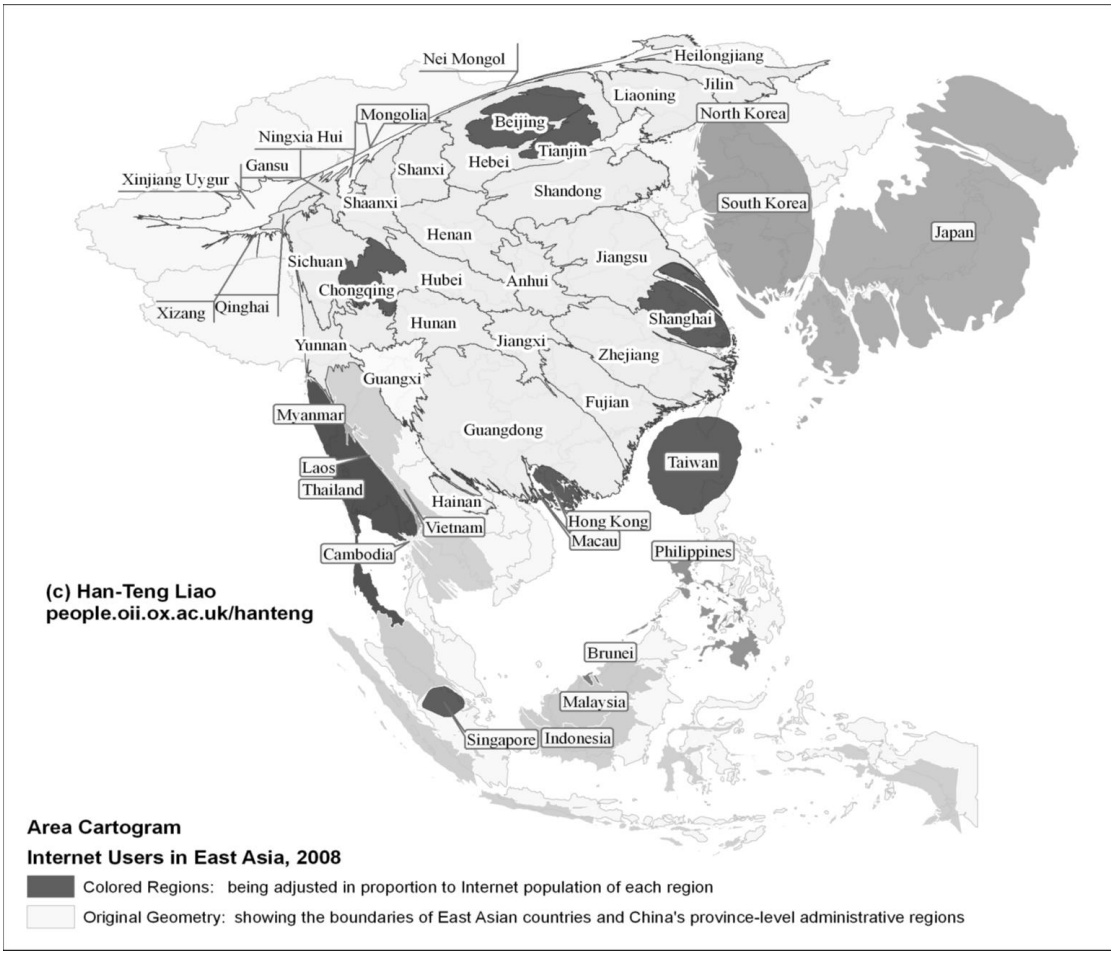

Figure 4: A cartogram of Internet users in 2008 East Asia

Network graph: affinity, core-peripheral, and diffusion

Another way to consider geo-linguistic factors is to draw a network graph, which aims to represent the interconnection between languages.[13] A preliminary network graph is presented in Figure 5, where each node represents a language version, with one-way and two-way directed links. The overall graph shows a core-peripheral pattern of the interconnections between different language versions of Wikipedia, with English (en), French (fr), German (de), Dutch (nl), Italian (it), Spanish (es), Russian (ru), Japanese (ja), and so on, constituting the core language networks. Additionally, some observations can be made about geo-linguistic affinity shown in this graph. For example, Chinese (zh), Japanese (ja), Korean (ko), and Vietnamese (vi) are depicted as fairly close to one another on the top-center of the graph; while Spanish (es), Portuguese (pt), and Catalan (ca) are shown with similar affinity at the right. At the bottom center of the graph are the Russian (ru), Slovak (sl), and Polish (pl). This seems to reflect the geo-linguistic affinity of East Asian languages (such as Chinese, Japanese, Korean, and Vietnamese) as well as Latin languages and Slavic languages that constitute some kind of a group. In addition, the three groups are relatively kept away from one another, with other core languages such as English (en) and German (de) in the middle. By considering both the overall core-

peripheral structure and the respective neighboring nodes, related hypotheses can be generated for further research, particularly with further data clustering tools.

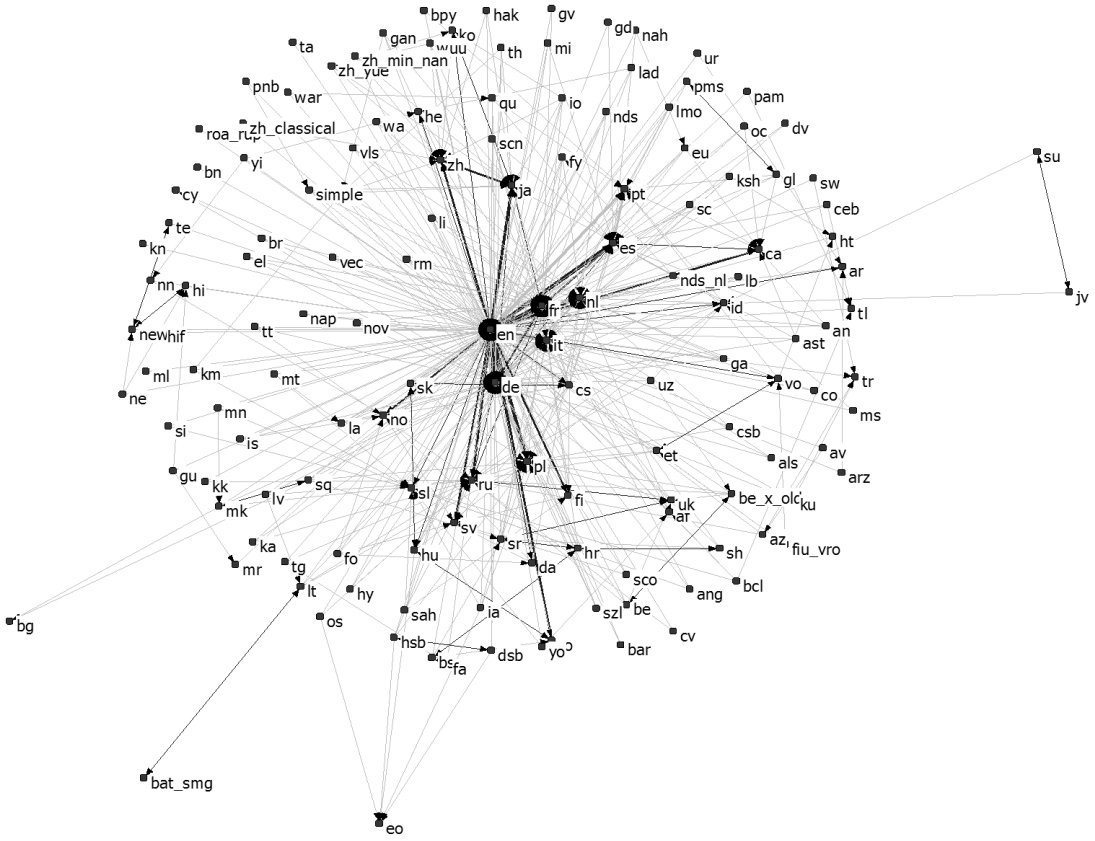

Figure 5: A network graph of selected language versions of Wikipedia

The core nodes such as English and German may be moved away from geo-linguistic conditions because many other language versions of Wikipedia connect to them, no matter how far they are geographically or linguistically located from core languages. It can then be hypothesized that if a language version is connected by many other versions regardless of geo-linguistic affinity, the node that represents that language will likely be in the centre of the network graph, suggesting that they constitute hubs amongst many nodes (Barabasi, 2003), or in other words generate a sense of universality. For a language version that is only connected by few other versions, the node that represents that language will likely be situated around the nodes that are geographically and/or linguistically close to one another. In fact, the examples of Javanese (jv) and Sundanese (su), shown in the far right of Figure 5, suggest not only their overall peripheral positions in the global networks of Wikipedia, but also their geo-linguistic affinity with one another. If the network graph shown in Figure 5 truly represents the interconnection pattern amongst Wikipedia language versions, then it cannot be claimed that geo-linguistic factors are overcome by ICTs such as the Internet and Wikipedia.

It cannot be claimed either that geo-linguistic affinity constrains every language. It is a much safer claim to make that geo-linguistic affinity is “distorted” somehow by ICTs in a

way that the core languages may attain universal "connectedness," just as the languages of French, German, Russian, Japanese, and even Chinese are not only connected to the core node of English, but are also getting closer to one another for that reason. Moreover, it somehow reflects the infrastructure realities of global networks as well. These language versions are brought closer to one another not only because of English, but also because of the ICT development. For example, whereas Javanese and Sundanese seem to be left out on the periphery (on the far right-hand side of Figure 5) both of them have to connect to the network of languages via Indonesian (id), which can be explained by their relationship to the Austronesian language family, and also reflects the hierarchy of international-national-regional languages, with Indonesian as a national language and Javanese and Sundanese as regional languages.

While the conjectures generated from the network graph shown in Figure 5 may be contestable, the purpose of the article is to show how geo-linguistic factors may or may not matter. In fact, the network graph of several language versions of Wikipedia has shown that geo-linguistic affinity may still matter, even without considering related geo-linguistic factors in making the graph. In other words, the geo-linguistic affinity is somehow reflected onto the network graph. The selected data used for generating Figure 5 is limited to those entries with less than three inter-language links, with the aim to highlight any affinity that can be captured by the spread of certain content in one language version to another. It should be noted that such affinity may be arbitrary for a specific entry. However, as the original dataset used for Figure 5 covers all inter-language links amongst all language versions of Wikipedia, such arbitrariness has been lowered by making sure there exists enough inter-language links for a link to appear in Figure 5. In other words, all the links that appear in
Figure 5 suggest that there exists a substantial number of links between the two nodes of language versions, while if no link appears between two nodes in Figure 5 it means that there exist no or some unsubstantial number of links between them. Thus, it is important to clearly define the construct of links and nodes in order to create a valid and meaningful network diagram. Again, further vigorous network analysis is required to confirm the conjectures of core-peripheral structure and clustering effects (how and why some languages are more likely to then become neighbors than others), which is beyond the scope of this article. Nevertheless, the results so far do suggest that geo-linguistic affinity does exist and that the relationship is not arbitrary. This can better be explained by briefly describing how the dataset underpinning the network graph is actually collected, framed, and constructed.

The data on which the network graph is based is a selected set of the inter-language links amongst all the language versions of Wikipedia. To Wikipedia readers, the inter-language links show up in a box with the title "languages" at the left column of almost every Wikipedia entry article, which contains a set of links that lead readers to other language versions of the same (or nearly equivalent) entry, as shown in Figure 6, where only four other language versions (German, French, Japanese, and Chinese) exist for the English entry of the "Green Dam Youth Escort."

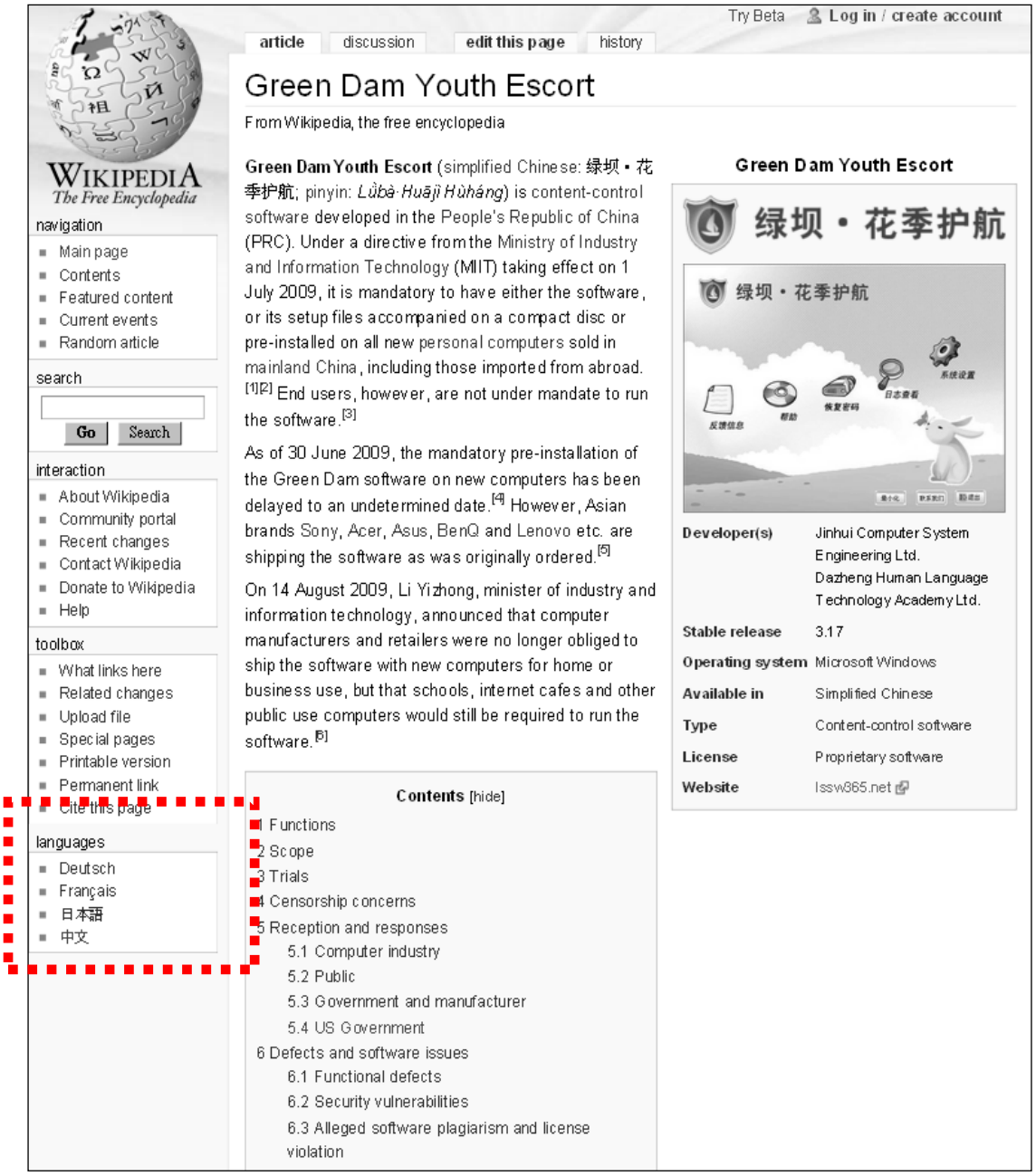

Figure 6: A typical Wikipedia entry page showing the location of the inter-language links

Such inter-language links can also be edited as part of the entry, with a straightforward syntax of the target language code followed by the title of the entry in that target language as shown below:

[[de:Green Dam Youth Escort]]
[[fr:Green Dam Youth Escort]]
[[ja:緑バ・ 花季護航]]
[[zh:綠壩·花季護航]]

Thus, from the example shown in Figure 6 and the syntax code, the English entry of the "Green Dam Youth Escort" is linking to four other language versions. Most of the time, it is assumed for an entry to grow along with the increasing number of inter-language links.

Certain popular (or universal) entries are destined to appear in almost all language versions such as Wikipedia itself. Some regional or parochial entries are expected to be bound within certain language versions. It is worth mentioning that many new entries in Wikipedia start as translation of articles that already exist in other language versions, a process which the inter-language links help to facilitate. Hence, it is possible for some parochial entries to spread across other languages, and the number of inter-language links for a given entry can be thought to grow from zero (meaning the entry does not link to any other language version), all the way to the sum of all language versions minus one (meaning the entry has links to every other language version that exists). The growing dynamic nature of inter-language links thus provides an observation site where the patterns of spread, diffusion, and distribution of links can be shown by drawing the corresponding network graphs. At any given moment, it is expected that some entries may have more inter-language links than others, with possible values from zero to near the sum of all language versions. It is expected that those that have fewer inter-language links can help to cluster languages into meaningful groups whereas those that have more inter-language links can be regarded as having universally popular content. Hence, although the sample is limited in scope (e.g., the entry shown in Figure 6 is excluded in our sample), it nonetheless highlights important ties between language versions.

<TX>Also, similar to the discussion on using cartograms for diffusion research in the previous section, several network diagrams can be created at different times so that a process of diffusion may be observed. The use of network graphs for diffusion research is expected to have several extra benefits that may not be offered using area cartograms. Firstly, because the network graph may show a core-peripheral structure, suggesting some hierarchical relationships, researchers may observe how the spread of inter-language links reinforces, reconstitutes, or shifts the existing hierarchical relationship. For example, as languages such as Japanese, Chinese, Russian, German etc. grow in terms of number of entries, will these languages reinforce the current central position of English? Will the central node shift from English to another language? Or, will these languages reconstitute in order to become central nodes of similar significance?

<TX>Secondly, unlike area cartograms where geographic affinity is already assumed and presented on the map, the diffusion patterns observed from network graphs reflect actual linking affinity, which in the case of inter-language links of the Wikipedia is likely to include geo-linguistic kinds of affinity. For example, it is conceivable that, borrowing from the concept of "geo-linguistic regions" (Albizu, 2007; Sinclair, 1996) for areas such as international and regional TV programming, some entries of popular culture may demonstrate how these cultural products may have been spread from one language domain to another. The dynamic dataset provided by the global Wikipedia project can thus be an important site of observation to empirically examine the geo-linguistic region concept, particularly as these inter-language links reflect actual developments online.

<TX> Thirdly, the use of network graphs can be regarded as an independent cross-check for the cartogram results because the geographic affinity is not assumed in the former while it is in the latter. For example, if some relationship that appears in the network graph cannot be explained by cartogram results, researchers will have to come up with explanations as to why such relationships may exist without clear geo-linguistic affinity. To sum up, conventional choropleth maps, cartograms, and network graphs can

show the static distribution and temporal diffusion of languages using clearly defined constructs as well as reliable data. In addition, the network graph can not only show possible core-peripheral structures but also provide an independent cross-check of geo-linguistic analytics drawn from cartogram results.

Discussion

Some may argue that geographic and linguistic factors do not matter to central nodes such as English because of the dominant position in Wikipedia—every other language version connects to it somehow. Still, if we take a longer historical perspective on how the global Wikipedia project expands and diffuses from the core language version to others, one argument can be made that geo-linguistic factors are rendered hidden as English becomes "universal." The "universality" of English as a core language depends on the fact that English has been the principal working language of the Wikipedia project (and the Internet) and that the English version of Wikipedia has been used as a central point of contact for inter-language links. However, with increasing numbers of entries in other language versions, it is expected that some entries, for various geo-linguistic reasons to be explored, may only exist in a few language versions. These entries, as has been shown in this article, can provide some important indication about other geographic and linguistic development patterns that may reinforce or challenge the core position of English. Although no conclusive claims can be made so far (even within the case of Wikipedia), this article has shown why geo-linguistic factors must be considered, where and how to examine major geo-linguistic developments online, and what influence this has on issues such as linguistic diversity and innovation diffusion.

Our analysis serves as a showcase of more comprehensive research to come. For example, our analysis provides some preliminary explanations about how languages are linked or connected on Wikipedia. However, as to "why" this is the case, more research is required. The main contribution of this article is to show how the geo-linguistic factors may be important and how researchers can capture them. We have also showcased the usefulness of geo-linguistic analysis by using two techniques, cartograms and network graphs. Thus, the empirical evidence we use is mainly to show the possibilities and potentials of such techniques (and their underlying propositions), not to argue for or against an empirical argument on specific cases.

The importance of geo-linguistic analysis is crucial in understanding the development, diffusion, and distribution of human knowledge online and offline. After all, no matter how we define knowledge, it has to be written and debated in a certain language, a process which somehow conditions (if not determines) the development, diffusion, and distribution of ideas. The truism can best be observed in the case of the Wikipedia project, where some core language versions (e.g. English) can be identified as major vehicles for "universal" knowledge. However, it should be noted that it remains to be seen for alternative patterns or paths of development, diffusion, and distribution of knowledge to occur across different languages. This is why it is essential for researchers to frame, collect, and analyze the geo-linguistic factors that address issues such as linguistic diversity and innovation diffusion.

Limitations and Future Development of Geo-linguistic Analysis

Observing the World Wide Web with specific techniques and theories that take into account the potentials and limitations of geo-linguistic factors, also requires an awareness of potential limitations. Cartograms, for example, are used as a visual method for approaching a specific problem. They must not be misunderstood, however, as direct visualizations as they "can be hard to interpret without additional information" (Fotheringham, Brunsdon, & Charlton, 2000, p. 26). They are deliberately exaggerated and can be perceived as unusually displayed. The benefits of cartograms become most obvious when they are applied widely and in ways that a variety of agents (individuals, governments, enterprises) can relate to. Tobler (Id.) paraphrased this as follows: "Satellite image globes have to some extent supplemented political globes and anamorphic globes might someday also be constructed" (p. 69).

This article demonstrates some of the possibilities of geo-linguistic analysis to show the benefits of analyzing how knowledge is developed, distributed, and diffused across languages. Two cutting-edge techniques (cartograms and network analysis) are deployed to explore the potentials of new ways of using geo-linguistic analysis. Nonetheless, the future expansion and refinement of such possibilities requires researchers to think thoroughly through the contexts and purposes of exercising geo-linguistic analysis online. The authors of this article strongly believe that research that uses and follows geo-linguistic data and analysis will contribute to various policy and research areas, for example, around the broader issues of information and communication technologies for development (ICT4D) as well as Internet governance. After all, they share the same concerns as to how certain geo-linguistic development patterns are made possible and rendered difficult. Finally, by showcasing two techniques applied to the global project of Wikipedia this article hopes to stimulate more methodical and conceptual experiments on using and rethinking this much neglected but essential component of Internet research and practice.

Notes

1. This article was first published in Araya, D., T. Houghton & Y. Breindl (Eds) Nexus: New Intersections in Internet Research, New York: Peter Lang, 55-75. It is republished here with the kind permission of Peter Lang Publishing Group.
2. One particular indicator for such developments are numeronyms like i18n (internationalization) and L10n (localization).
3. For more updated numbers and a list of the local domains and languages see http://www.google.com/language_tools?hl=EN (last accessed February 15, 2010)
4. For more updated numbers and a list of the Wikipedias see http://meta.wikimedia.org/wiki/List_of_Wikipedias (last accessed February 15, 2010)
5. Wikimedia defines regions and region codes as follows: 'Regions are parts of the world where the language is spoken in substantial amounts (compared to total number of speakers). Regions where a language gained presence only by a recent diaspora are generally not included. Region codes: AF:Africa, AS:Asia, EU:Europe, NA:North America,

OC:Oceania, SA:South America, W:World Wide, CL:Constructed Language' (cf. http://stats.wikimedia.org/EN/Sitemap.htm, last accessed February 15, 2010)

6. For Wikimedia's language proposal policy see http://meta.wikimedia.org/wiki/Meta:Language_proposal_policy (last accessed February 15, 2010)

7. This means it must be listed in an ISO-639 database, or standards organizations must be convinced to create an ISO-639 code for a 'new' language.

8. This, in most cases, excludes regional dialects and different written forms of the same language.

9. This requirement, which must be met for the final approval, is discussed in an open discussion. To do so, a project will be initiated where interest by individual speakers or supporters of the language is registered and arguments for and against the admission of the new language are gathered. Then a decision will be made by the language committee.

10. Cf. http://stats.wikimedia.org/wikimedia/animations/growth/index.html (last accessed February 15, 2010)

11. One useful tool for analyzing Wikipedia's statistics is the Wikimedia Toolserver which has provided necessary assistance for the relevant materials used in this article. Our particular thanks are due to Daniel Kinzler at Wikimedia Deutschland. Wikimedia Toolserver is hosted by Wikimedia Deutschland with the assistance of the Wikipedia Foundation, and provides sophisticated access to the Wikipedia database for researchers. For more information see https://wiki.toolserver.org/view/Main_Page (last accessed February 20, 2010)

12. Cf. http://www.worldmapper.org/ (last accessed February 15, 2010)

13. For a brief methodological note, the selected raw data was generated with the Wikimedia Toolserver, which was then processed with the programming scripts written by Han-Teng Liao to produce a network graph file. In a next step, the network graph file was fed into social network analysis and exploring tools such as NodeXL and UCINET. The tentative graph shown in this paper is produced by UCINET, with the spring embedding layout. The settings for the layout are as follows: the criteria is based on "Distances + Node Repulsion"; the starting positions "Gower scaling"; the number of iterations 30; the distance between components 30; the proximities "geodesic distances."

References

Albizu, J. A. (2007). Geolinguistic regions and diasporas in the age of satellite television. *International Communication Gazette*, *69*(3), 239–261. doi: 10.1177/1748048507076578

Barabasi, A. L. (2003). *Linked*. New York: Penguin.

Bruce, S., & Yearley, S. (2006). Diffusion of innovation. In *The Sage dictionary of sociology* (p. 73). London: Sage.

Cartwright, D. (2006). Geolinguistic analysis in language policy. In T. Ricento (Ed.), *An introduction to language policy* (pp. 194–209). Malden, MA: Wiley-Blackwell.

Fotheringham, A., Brunsdon, C., & M. Charlton (2000). *Quantitative geography*. London: Sage.

Gastner, M. T., & Newman, M. E. J. (2004). Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(20), 7499–7504. doi: 10.1073/pnas.0400280101.

Gerrand, P. (2007). Estimating linguistic diversity on the Internet: A taxonomy to avoid pitfalls and paradoxes. *Journal of Computer-Mediated Communication, 12*(4), 1298–1320, doi: 10.1111/j.1083-6101.2007.00374

Hecht, B. & Gergle, D. (2010), *The Tower of Babel Meets Web 2.0*, CHI2010, April 10–15, Atlanta, Georgia.

Liao, H. (2009). Conflict and Consensus in the Chinese Version of Wikipedia. *IEEE Technology and Society Magazine*. Retrieved March 30, 2009, from http://www.ieeessit.org/technology_and_society/default.asp

Paolillo, J.C. (2007). How much multilingualism on the Internet? Language diversity on the Internet. In B. Danet & S.C. Herring (Eds.). *The Multilingual Internet* (pp. 408–430). Oxford, UK: Oxford University Press.

Petzold, T. (2010). *36 Million Language Pairs: Generative Multilingualism in Digitally-Enabled Societies (Discussion Paper for the Twelfth Berlin Roundtables on Transnationality).* Berlin: Social Science Research Centre. Retrieved March 10, 2010, from http://www.irmgard-coninx-stiftung.de/fileadmin/user_upload/pdf/Cultural_Pluralism/Language/Essay.Petzold.new.pdf

Raisz, E. (1938). *General cartography*. New York: McGraw-Hill.

Sinclair, J. (1996). Culture and trade: Some theoretical and practical considerations. In E. G. McAnany & K. T. Wilkinson (Eds.), *Mass media and free trade* (p. 444). Austin, TX: University of Texas Press.

Tobler, W. (2004). Thirty-five years of computer cartograms. *Annals of the Association of American Geographers, 94*(1), 58–73, doi**:** 10.1111/j.1467-8306.2004.09401004.x

Van der Merwe, I. (1993). A conceptual home for geolinguistics: Implications for language mapping in South Africa. In Y. J. D. Peeters & C. H. Williams (Eds.), *The cartographic representation of linguistic data (Discussion Papers in Geolinguistics, Nos. 19-21)* (pp. 21–33). Stoke-on-Trent, UK: Staffordshire University.

Wikimedia (2009). Emerging strategic priorities. Retrieved February 2, 2010 from http://strategy.wikimedia.org/wiki/Emerging_strategic_priorities