

Title of Paper

Information Poverty due to Linguistic Diversity in South Asia and Sub-Saharan Africa:

Policy-recommendations for making online content available in local languages

Author Details: Author 1: Ranjit Goswami, Associate Professor, IIFT Kolkata,
Ranjit.goswami@gmail.com

Author 2: Dr. S K De, Associate Professor, IIT Kharagpur, drskde@vgsom.iitkgp.ernet.in

Author 3: Dr. B. Datta, Assistant Professor, IIT Kharagpur, bd@vgsom.iitkgp.ernet.in

Contact Details: Vinod Gupta School of Management, Indian Institute of Technology, Kharagpur –
721302, www.som.iitkgp.ernet.in, Ph. No. - +91-3222-282295, Fax - +91-3222-255303

For correspondence: Please contact 1st author, Ranjit Goswami at his e-mail id

Abstract

We start with a fundamental question that drives the actions of most local and global policy-makers: out of two categories of people (not mutually exclusive), the *'have nots'* and the *'know nots'*, which one is more difficult to eradicate (one may pose the question differently – solving which of these two problems is likely to solve the other). A lot of attention and resources have been deployed and are committed on the challenge of uplifting the *'have nots'* to the section of bare minimum *'haves'* category. A cause and effect study between these two sections of people essentially show a mutual dependency that eventually lead to a vicious cycle of poverty to information poverty to back again poverty which has historically been difficult to eradicate, and studies have often established education and access to timely information to be a sustainable remedy to both these perpetual problems. The role that Internet can play in this background towards empowering the billions of impoverished across two of the most underdeveloped regions, namely South Asia (SA) and Sub-Saharan Africa (SSA), home to world's largest number of illiterates and poor people (70% or even more together) is immense. With anywhere, anytime accessibility of rich Internet content aided by its falling prices and increased connectivity and easy to use features, to parts of rural and even to inaccessible remote areas; online content can effectively act as a low-cost feasible solution not only to provide basic education, but also to deliver meaningful information and content to the millions of primary-level educated people within the underprivileged sections of SA & SSA, thereby enabling them to integrate and exploit various socio-economic opportunities arising from economic developments.

However rich linguistic diversity, both in SA and SSA, poses a challenge to that opportunity. We argue that content development to information access to literacy, all leading to socio-economic developments, do face additional difficulties arising from linguistic diversity for SA and SSA, regions already plagued with low level of content generation and access characteristics in local languages. A closer examination of the 'growing ocean' of online content reveals that SA scores poorly in local language content development where English is primarily used for Internet usage, though nearly 90% of people of India (major country of SA, with other nations following similar trends) don't use English as its language (including 2nd and 3rd languages). For SSA, we qualitatively examine whether linguistic diversity indeed has any negative correlations with Gross National Income (GNI) and Internet penetration; and qualitatively find they indeed

are inversely related in 80% or more cases. We also examined one case (globalvoicesonline.org) that focused on development of local language content, critical to reap benefits from content for development for SA and SSA, but found it to be inadequate both in content and in representation in relevant local languages in proportion to the severity and scale of the problem in SA and SSA. We conclude with alarming note that unless war-footing action is adopted to generate local language content for content-deficient languages in online environment (or effectively supported by software like Google Translation or local language supports for User Generated Contents like Google Blogs in Hindi) in the linguistically diverse backward regions of our world, much of the benefits that can be derived from increased reach of freely available online content would be lost, causing an escalation of information poverty to the 'bottom of the pyramid' section of people in South Asia and Sub-Saharan Africa.

Keywords

Online content, native language, information poverty, growing inequality, Linguistic diversity, South Asia, Sub-Saharan Africa, medium of language for accessing online content, online content as a social tool

Introduction – Information inequality at play in South Asia and Sub-Saharan Africa

Internet and its broader Information and Communication Technologies (ICTs) with their disruptive characteristics have offered many leapfrogging opportunities to the developing world. Data, information, knowledge or wisdom; they all so long were in the private domain of the rich, or the wise or working professionals. Rarely had the information-poor section of the society, who needed a fraction of that accessibility to that pool of information, had timely access of it. ICTs, more with the revolutionizing spread of online content, have killed that myth.

Lyman (2000)¹, in his well-known Berkeley project titled '*How much information*' in the beginning of the knowledge-millennium showed that the total yearly production of new information in our world amounts to about 250 megabytes/person/year. In a further revised estimate in 2002, the same study showed that the amount of information production doubled to more than 800 megabytes/person/year, equivalent to 30 feet

¹ Lyman, P 'How Much Information' <http://www2.sims.berkeley.edu/research/projects/how-much-info/internet.html>

of books.² According to Lloyd³, his theory of information being a physical entity gets stored in the 'quantum computer' of the universe. In this way, he tried to answer questions like '*how much information do we have in the universe*' in his book, backed with astronomical figures of 10^{120} bits. These only prove that there is no dearth of relevant meaningful information that the underprivileged needs most from the dimension of availability. What needs to be explored is accessibility and understandability part of that information for the target segments.

The critics of the Berkeley study therefore rightfully felt that information is not comparable to a static physical property like mass. Information is an active thing, defined by its creation and use. It is more like velocity or acceleration. Information or for that matter content therefore serves its purpose better when it's meaningfully used by more number of people. This essentially can happen in two ways (1) Certain categories of content which apparently appears to be repeated and run-of-the-mill to the information-rich users; can appear new to the information-poor segment of population, and therefore be more productive, and (2) the content which focuses on specific local communities are often found to be more useful locally, productively again. We kept extremely innovative original global content (invention/discovery type) beyond our purview of examples, while looking at the effectiveness of content based on its usage and its expected socio-economic impact.

With the rapid growth in Web 2.0 applications with YouTube generations and amidst the boom in the User Generated Contents (UGC), content now can be generated by anybody, anywhere and at any point of time. The measure of online content should therefore be more in its applications, and suitability for the segment who had least access of it so long. In the article '*How do we make all this information more useful?*', BusinessWeek⁴ cited the example of Galileo, who understood the importance of access to information when he published his works in Italian, the language of the masses, to share his work with

² Roy, Williams 'Data powers of ten' accessed from <http://www2.sims.berkeley.edu/research/projects/how-much-info/datapowers.html>

³ Lloyd, Seth 'Programming the Universe' <http://www.amazon.com/Programming-Universe-Quantum-Computer-Scientist/dp/1400040922>

⁴ BusinessWeek 'Using What You Know' April 21, 2006 http://www.businessweek.com/technology/content/apr2006/tc20060421_260198.htm?campaign_id=hp_views&campaign_creative=Paul%20Horn

populace rather than in Latin which was the language of the intelligentsia in those days. A similar trend is observed in present times with local languages of masses from SA and SSA against leading Internet languages (English, and others), the language of the intelligentsia in these two regions.

In this paper, we examined how much of the booming content is actually relevant, purely from languages perspectives, for the populace of the two of the most backward regions of the world, constituting nearly 40% of the world population. Communities having access to timely information access does play a key role in the development process. We wanted to check whether linguistic diversity does create additional hurdles in enabling information access from existing online contents to the people of South Asia (SA) and Sub-Saharan Africa (SSA). Presently these two poorest regions⁵ of the world also have the lowest access to information and communication technologies. The gap in access to ICTs, the so-called 'digital divide' often follows and reinforces the existing inequality and poverty patterns. Online content, if it has to prove its utility, SA and SSA would be its perfect testing ground. True, the scenario becomes more relevant when online content is accessible to many more in SA and SSA, however we believe that's now a question of time, that too in years and not in decades.

We did observe strong linkages between linguistic diversity with economic backwardness (Per Capita GNI where these two regions rank lowest)⁶. Levels of Internet penetrations to per capita GNI in SSA for its ten most linguistically diverse nations was worse than average for SSA itself. Availability of local language content in SA also scored poorly for local issues compared to English. Moreover, in line with our anticipation, we found that linguistic diversity does pose an additional socio-economic hurdle and thereby hampers information flow in SSA⁷. Eight of the ten nations had lower than average GNI (lowest being

⁵ Pigato, Miria 'Information and Communication Technology, Poverty, and Development in sub-Saharan Africa and South Asia' 2001 <http://www1.worldbank.org/wbiep/decentralization/library1/pigato.pdf> found that 70% of the 1.2 billion poor people living below a dollar a day in the world lived in SA and SSA.

⁶ South Asia had a per-capita GNI of \$ 766 whereas Sub-Saharan Africa had it \$842 in US dollars; and in PPP Sub-Saharan Africa had per-capita GNI of \$2031 against that of \$3443 for South Asia. Source: GNI Per Capita, 2006 Atlas method and PPP, <http://siteresources.worldbank.org/DATASTATISTICS/Resources/GNIPC.pdf> accessed on 30th August 2007

⁷ Ten most linguistically diverse countries in sub-Saharan Africa << Sociolingo's Africa <http://sociolingo.wordpress.com/2007/04/18/ten-most-linguistically-diverse-countries-in-sub-saharan-africa/> accessed on 30th August 2007. Source of information is population figures - UNPD 2001; language information &

1/3rd of SSA) and nine of these ten nations had lower than average Internet penetration⁸ (for Africa, even adjusting for South Africa's gross overweight position here⁹). For SA, we applied an indicative measure, using a newly-launched-application ('Indic on-screen key boards'¹⁰, introduced by Google keyword search) in Bengali and Hindi to compare the online content generation (keywords being local cities from SA) in local languages against the content in English for the same keyword. Incidentally, Hindi and Bengali are two of the highest spoken South Asian languages. The content in the local languages (Hindi and Bengali) turned out to be at best 1.3% to a worst of 0.002%, compared to the content of English.

Linguistic diversity and backwardness in SA and SSA amidst English dominated online content

During the last couple of centuries, both SA and SSA shared many socio-economic characteristics. Another historical commonality have been the colonial rule, and legacy of colonial languages over a linguistically disparate local population. Linguistic diversity, a rich cultural heritage as seen in developed countries as in Europe, can also pose to be an additional challenge in regions like SA and SSA, when it comes with illiteracy, poor quality of primary education and poverty¹¹. These two regions faced additional complexities in its effort of educating its linguistically diverse populace with relevant timely information for decades as no mass-media channels of past could support this gigantic exercise cost-effectively, primarily due to limitations associated with mass-media channels of past. With the opportunities the new-age media, online content generation or its conversion from leading Internet languages into major local languages on a dynamic basis may not be as daunting a proposition for SA and SSA, due to the various advantages online media offer. However, till date very less has been achieved on making online content

GDI - Ethnologue 2000 CDROM

⁸ Internetworldstats.org 'Africa Internet Usage and Population Statistics' accessed on 28th August, 2007 (<http://www.internetworldstats.com/stats1.htm#africa>)

⁹ South Africa alone accounted for almost one-third of all African Internet connection in 2000 ('For Most Africans, Internet Access Is Little More Than a Pipe Dream, <http://www.ojr.org/ojr/workplace/1079109268.php>)

¹⁰ Google Indic On-Screen Key Boards used on 23rd August 2007 for Bengali and Hindi as authors were familiar with these two languages, other than English. <http://labs.google.co.in/indic.html>. Media coverage showed that the Google Indic On-Screen Key Boards application itself was introduced in August 2007 helping user type keywords in Google search using local Indian languages even from an English-language keyboard.

¹¹ National Geography: Lesson Plans – African Language Diversity 'The linguistic diversity of Africa is considered by some to be a problem for its people.' <http://www.nationalgeographic.com/xpeditions/lessons/18/g912/afrolanguage.html> accessed on 30th August 2007.

available in major local languages in SA and SSA for their populace. The alternative solution being educating their populace in another leading Internet (foreign) language with ample content, more challenging, this two-fold approach of converting foreign-language content, relevant for local people, in local languages aided with another focused approach to generate (at least in one leading local language and then converting in other locally relevant languages) local content where global content is inadequate seems to be viable option.

There are reasons to be optimistic with this disruptive social face of the Internet when one finds that the number of people accessing the Internet from SA and SSA regions are growing fast (usage growth of Internet from Africa and Asia are highest for the period of 2000-2007, at 643% and 282% respectively¹², true the base was lowest). The access cost of internet is dropping fast and more and more content producers increasingly adopted free-access policy (ad-model against paid-subscription model). Most encouraging is the quantum of information and online content – relevant from local issues to global ones that already exists online, and also gets created daily on a regular basis. However, the challenges as faced by the linguistically diverse underprivileged societies in SA and SSA come from another dimension and that is the medium of the online content. Only a small fraction (around 10%) of the population in these two most backward regions of the world can follow a language in which Internet content has mostly been generated. Internet contains insignificant content in the native languages of SA and SSA (or can even read as 2nd or 3rd languages). It effectively leads to a situation of *'Information, Information everywhere'* for the people, who comprehends any of the leading Internet languages (primarily English) leading to 'too much information' syndrome in one hand, whereas for billions of people, who don't comprehend any of the leading languages of Internet and need information the most, find themselves in opposite situations. It is understandable that languages with poor content further races to the bottom in content generation as people, who can contribute content in that local language migrates to the alternative global language for better prosperity, threatening the survival of that language in the long run¹³.

¹² op.cit. <http://www.internetworldstats.com/stats.htm>

¹³ Stephen, A 'The Atlas of the world's languages in danger of disappearing'

Considering the 'haves' and 'have nots' of the society, we observe conspicuous consumerism in a small insignificant section of the society in contrast with a lack of basic infrastructure and pathetically low level of consumption for the other significant section, raising the specter of rising inequality globally. There is a more severe dimension of information inequality emerging with increasing dominance of Internet as a media, and again by dominance of English (or other European languages) language in the online environment. Information inequality, though as a concept difficult to measure, if could be measured effectively would probably show a much worse situation than the economic inequality in which we live on present world.

Though there has been a rapid growth in telecom and Internet access over the mobile-network, mobile content still effectively remains a miniscule of information/content access over desktop environment for usage of online content. True, developments in mobile applications are encouraging; and content can eventually be accessed by any device – mobile or desktop. The question therefore, irrespective of device, is having content in local languages as eventually devices that can access online content grow all over, in SA and SSA as well. In India, most mobile keypads and computer keyboards are in English, and much of the locally relevant online content also happen to be in English. And this scenario is predominantly true for the other nations of SA and also for SSA. So this effectively leads to a situation in present times, where a person with bare minimum education and connectivity can't access the content relevant for him because of the language barrier. However, the above aspect is quite complex and still happens to be an evolving area in which it is difficult to conclude anything with clarity. In 2006¹⁴, Yu in an extensive literature review over information inequality observed multifarious perspectives and standpoints with divergent and contradictory views of researchers. Even the local language content was quite confusing for any researcher to draw any firm conclusions.

http://portal.unesco.org/education/en/ev.php-URL_ID=13144&URL_DO=DO_TOPIC&URL_SECTION=201.html

¹³ http://en.wikipedia.org/wiki/Swahili_language

¹⁴ Yu, Liangzhi 'Understanding information inequality: Making sense of the literature of the information and digital divides' *Journal of Librarianship and Information Science*, Vol. 38, No. 4, 229-252 (2006)
<http://lis.sagepub.com/cgi/content/abstract/38/4/229>

Barring China, Japan, South Korea and parts of Arab world¹⁵, this problem is predominant for all non-European languages. Obvious contributions in the Internet growth and its usage in many of these countries (China, Japan, South Korea; and none of these nations faced language diversity as SA and SSA faces) can be linked with the local language content developments. Nations from SA and SSA have adopted foreign languages as national language, more for the elite, and therefore content generation in diverse local languages have suffered.

China overtook the U.S. in telecom user base, and it is expected that China would soon repeat that feat with Internet user base as well¹⁶. Unlike China, Japan, South Korea, and Arab World, where we may be witnessing ample online content development in local non-European languages lately; SA and SSA disturbingly stands out as exceptions in developing the local language content for its populace. The root problem may again lie in the linguistic diversity in these two regions, unlike any other regions. When we keep in mind the very fact that the Internet is growing at an annualized rate of 18% and crossed a billion users in 2005 and a second billion users is likely to follow suit by 2015, bringing a dramatic change in its worldwide usability; this shortfall of online content in local languages is likely to be bigger and bigger hurdle for effective deployment of web as a social tool in these two parts of the world. It becomes imperative that the second billion users points to the growing interactive media's compelling value¹⁷, and unfortunately underprivileged and the large majority of people from SA and SSA would not be a part of that interactive value, due to reasons already documented in literature of digital divide, but also due to the language barrier where the digital barrier could be overcome, but not the medium of content as it stands in present times.

¹⁵ The only four languages that feature in Top ten languages used in the web (<http://www.internetworldstats.com/stats7.htm>, based on usage) and Internet Statistics: Distribution of languages on the Internet, Chart of Web content (millions of webpages by language) 2002 (<http://www.netz-tipp.de/languages.html>, based on content categorization in different languages) which again includes Japanese, Chinese and Korean in top 11 languages as per pages available.

¹⁶ Forbes.com 'China surpasses U.S. in Internet Use' dated 04.03.06 at http://www.forbes.com/2006/03/31/china-internet-usage-cx_nwp_0403china.html. The article talked about two different estimates – as per official The China Internet Network Information Center (CNNIC) estimate it still lags when it comes to numbers of users. However as per another estimate by Dr. Zhang, Chairman and CEO of Sohu.com, in terms of Internet usage (and even in number), China may already be much ahead.

¹⁷ 'One billion Internet Users' http://www.useit.com/alertbox/internet_growth.html

Content development is a race where more one lags behind, more one becomes the loser because better-offs either adopt the language with adequate content to remain competitive or people lose out the opportunities. Without much focus and incentives to develop content in local languages of SA and SSA, these two worse-off regions are likely to face two distinct challenges in terms of leveraging this tremendous reach, quality and quantity of knowledge economy as (1) Internet growth rate itself may slow down (number of online users); or (2) Usage of Internet slows down due to lack of local language content. The edge of digital learning opportunity may therefore be lost in both SA and SSA where it's needed most in our world.

As we moved to examine this inequality of content development in different languages, we had a look at several of the estimates that gave us various measures on the size of the online content. Estimations of 2000 did put a figure of 2.5 billion documents accessible to all in 'surface web'¹⁸, whereas 'deep web' was estimated to be 400-550 times larger than the 'surface web' with daily additions of 7.3 million pages per day. Comparable figures of 2007 as per Internet Archive¹⁹ are 85 billion web-pages (including sites no longer live). Other estimates varied between 15 to 30 billion web-pages, the shift being towards the higher side.

Going by the Internet usage latest data for 2007²⁰, we roughly get 25 web-pages per internet user in 2007, whereas comparable figure for 2000 and 2005 was respectively at 7 and 20 web-pages/user respectively²¹. This clearly demonstrates that the number of web-pages over Internet is growing at a much faster rate than the users. However, when we examine the language in these web-pages, we find

¹⁸ Definition of: surface Web 'Content on the Web that is found in search engine results' (source: http://www.pcmag.com/encyclopedia_term/0,2542,t=surface+Web&i=52273,00.asp) whereas 'Deep web' is defined as 'Content on the Web that is not found in most search engine results, because it is stored in a database rather than on HTML pages. Viewing such content is accomplished by going to the Web site's search page and typing in specific queries. LexiBot was the first search engine to actually make individual queries to each searchable database that it finds. Deep Web also includes password-protected content on the Web available only to members and subscribers. (Source: http://www.pcmag.com/encyclopedia_term/0,2542,t=deep+Web&i=41069,00.asp)

¹⁹ Internet Archive (IA) – Wayback Machine <http://www.archive.org/web/web.php> . As per IA, 85 billion web-pages have been archived since 1996 and therefore does not include prior period pages as on 22nd August' 2007

²⁰ World Internet Usage Statistics <http://www.internetworldstats.com/stats.htm> accessed on 22nd August 2007

²¹ Internet growth 2000 to 2005 <http://www.internetworldstats.com/pr/edi008.htm> accessed on 22nd August 2007

that out of 25 web-pages/user (in 2007), 14 (netz-tipp-de, 2002) to 20 web-pages (ITU, 1999)²² are in English. That's the tip of the iceberg, for SA and SSA at least.

Wikipedia, the online encyclopedia edited and generated again by netizens all over showcasing the democratic characteristics of Internet as a media, contained some 8.2 million articles in 253 languages against Encyclopedia Britannica containing some 500,000 articles²³. An elementary language analysis of Wikipedia showed that there were only Telegu, Nepalese, Bengali, Hindi, Marathi and Tamil from SA languages (in that order) within the top 70 languages in which Wikipedia content was generated and made available (on 4th September, 2007 each of these languages having 10,000+ articles in respective languages). Swahili was ranked at 88th level with only 5863 articles in that language. And Wikipedia definitely points out that when netizens themselves created web-pages independently; more content was created in other local languages as dominance of English in Wikipedia was much less at 24%²⁴ compared to higher measures when all of Internet is considered.

When it comes to language, back in 2000, 78% of all web-pages were in English though 50% of all internet users back then were native English speakers²⁵. Moreover, Cyber world has no borders and knows no boundaries. However, research has shown that venturing out of one's locality may be driven by the need to accomplish a task or to fulfill a goal or simply to satisfy one's curiosity. Wang and Servaes (2000)²⁶ pointed out that, the importance, significance and relevance of the global are not as great as that of the local.

South Asia, comprising primarily seven nations²⁷ i.e. Bangladesh, Bhutan, India, Maldives, Nepal,

²² International Telecommunications Union, 1999. "Challenges to the Network: Internet for Development."

²³ Rushe, Dominic 'Online encyclopedia aims to roll over Google' accesses from 'Today's Zaman' dated 2nd September, 2007 <http://www.todayszaman.com/tz-web/detaylor.do?load=detay&link=121066>

²⁴ Kist of Wikipedia showed a total of 8349480 articles as on 4th September 2007 against English language articles of 1987632. http://meta.wikimedia.org/wiki/List_of_Wikipedias. Accessed on 4th September, 2007.

²⁵ op.cit. 'How Much Information?'

²⁶ Wang, G. and Servaes, J. & A. Goonasekera 'The new communications landscape: Demystifying media globalization' (pp.1-18), London: Routledge.

²⁷ As per CIA World Factbook entries on South Asia (<http://www.columbia.edu/cu/lweb/indiv/southasia/cuv1/Fact.html>). However United Nations further includes

Pakistan and Sri Lanka, ranks highest in the world's most densely-populated regions (315 people/sq. km, seven times world average). The mother tongue of the vast majority people in this region is not English (though English is the official language of government and that of business world). Moreover, the online content developed in the local languages (numbering 26 major ones, if not more; please refer Image 1) is abysmally poor, compared to the population figure in each of these major native languages.

A surprising similarity is observed in the other backward regions of the world in Sub-Saharan Africa that comprises 42 African countries of mainland Africa (and 6 island nations), having a population of nearly 788 million. Linguistic diversity in SSA demonstrates similar characteristics of South Asia, where language of the masses are mostly different from the official language of the government/business world. Invariably one finds economic and social backwardness to be one of the highest in these two regions. As per a Food and Agriculture Organization (FAO) 2006 report²⁸, there still remains 820 million chronically hungry people in developing countries, with nearly half of that numbers originating in South Asia and one-third in Sub-Saharan Africa (balance 17% 'pockets' of hunger is in Latin America and in rest of Asia).

The above background makes one obviously ponder on any positive linkages between diversity in languages and lack of prosperity, though a look at Europe clearly defies that. When one further takes into consideration that Europe was the pioneer in industrial revolution and happens to be one of the most developed parts today, socio-economically, one clearly perceives linguistic diversity in Europe to be different from that of SA and SSA. Due to colonial legacy of most European nations, when one adds native speakers with secondary speakers, most European languages rank high in terms of global usage compared to domestic native speakers. Therefore, we find that six of the top ten Internet languages are of European origin (English, Spanish, French, German, Portuguese and Italian) accounting for a whopping 57% of the language of Internet population.

Afghanistan and Iran within Southern Asia (<http://millenniumindicators.un.org/unsd/methods/m49/m49regin.htm#asia>). Certain other classifications of South Asia does further include Myanmar and Tibet (http://en.wikipedia.org/wiki/South_Asia). We have excluded British Indian Ocean Territory from South Asia though logically (and as per UN), its part of South Asia. Accessed on 29th August, 2007

²⁸ The hunger project – frequently asked questions - <http://www.thp.org/faq.html> accessed on 30th August, 2007

UNESCO (2006) studies²⁹ stated that less than 60% of the total adult population in South Asia (and West Asia) and Sub-Saharan Africa can barely read and write. It has one of the lowest adult literacy rates in the world. The picture of the early-stage school drop-outs is equally shocking. Of these two regions, the same UNESCO studies pointed out that in half of South (and West Asian countries), of a cohort of pupils who enroll in primary education, less than 65% reached the last grade, whereas for Sub-Saharan Africa, the figure was at 60%. In such stark ground realities, relevant content in other languages, even if made available and accessible free of cost anywhere and at anytime, effectively means nothing for the vast populace of these two regions.

We did not find any focused study that ever tried to link the linguistic diversity, which could also be one of the major root sociological problems when it comes to providing education and making information available and meaningfully accessible (media reach) to underprivileged linguistically diverse people. We believe that linguistic diversity poses a serious challenge to content management in local languages for SA and SSA. Linguistic diversity creates challenges for primary education, information flow across nation in one side to social and economic development on the other to even cultural and national ones; as people divide in smaller groups in line with languages than following mainstream national identities.

ICT4D with media convergence: online content as a social tool faces the language hurdle

Highlighting the enormity of the problem of reaching the *'have nothing'* section of people, Wagner (2001)³⁰ stated that: *'Most policy makers, researchers and practitioners could at least agree on one thing: Reaching the poorest of the poor was (is) going to be the most difficult of challenges'*. And though economically the poorest of the poor mostly happen to be the poorest of the poor in other social

²⁹ EFA Global Monitoring report 2006 – Literacy for Life – Regional Overview – for South Asia (<http://unesdoc.unesco.org/images/0014/001497/149783E.pdf>) and for Sub-Saharan Africa (http://www.unesco.org/education/GMR2006/full/africa_eng.pdf). Another study by Oyeyinka & Lal (part of UN University study) stated average adult literacy in SSA to be around 49% compared to 81% for other developing countries with similar enrollment ratios in schools (The Internet diffusion in Sub-Saharan Africa: A cross country analysis <http://www.intech.unu.edu/publications/discussion-papers/2003-5.pdf>)

³⁰ Wagner, Daniel A, 'IT & Education for the Poorest of the Poor: Constraints, possibilities and Principles' TechKnowLogia, July/August, 2001, http://www.techknowlogia.org/TKL_active_pages2/CurrentArticles/main.asp?IssueNumber=12&FileType=PDF&ArticleID=304

measures as well; breaking this vicious chain of poverty cycle with the help of information and communication technologies (ICTs) offered an opportunity. However, while examining how to convert the possibility into reality, one realizes that making relevant information available to the significant number of information-poor section of the society remains the biggest and most difficult challenge for the policy-makers, as other physical concerns on accessibility are getting addressed due to innovations, market forces and also better policies. Much of the exhaustive studies in this field focused on different dimensions of availability (device, connectivity), accessibility (price-points) and skill set (how to use computer/internet/mobile devices. Search engines are probably easiest to use over desktop, and are most useful for any). Most of these are taken care of with better reach, falling prices and user-friendliness of the products/applications. The softer issue focusing on the dimension of language as a barrier to online content has never been studied in the context of these two regions. And this can potentially derail the advantages that could have been drawn from the improved and improving physical reach as the mind of any user still remains aloof due to language barrier that the physical reach brings (or brought in so far).

In 2003, the United Nations' World Summit on the Information Society (WSIS) officially acknowledged for the first time the emerging importance of ICTs as the main impetus for the developing nations to achieve economic growth (ICTs for Development - ICT4D)³¹. However, its prioritization didn't include content availability for the local communities in local languages; it rather focused on the other physical parts of the connectivity to get the penetration rolling in the form of hard infrastructures. Thas (2005)³² felt that WSIS failed to acknowledge the critical importance of integral information and communication to the development of the human society. We strongly believe that ICTs should be examined with their convergence abilities, and not in isolation from media. Although an unprecedented media convergence has already taken place in the developed world driven by the development in the ICTs, we feel that convergence of media and ICTs has been much less in regions like SA and SSA, both in terms of reach and content availability in local languages. However, ICTs single-handedly offered the leapfrogging

³¹ <http://www.itu.int/wsis/index.html>

³² Thas, Angela M K 'PADDLING IN CIRCLES WHILE THE WATERS RISE: GENDER ISSUES IN ICTS AND POVERTY REDUCTION' (2005) HTTP://WWW.GENDERAWARDS.NET/THE_AWARDS/2005A/MEDIABRIEF.HTM

opportunity of quality media reach in the form of convergence to the information-deficit section of the populace, which can potentially be biggest catalyst for the much-needed upliftment of underprivileged sections in SA and SSA at unbelievably low costs (both at investment and operational levels) and anytime, anywhere accessibility features.

So there was optimism with disruptive potential of ICTs in tackling many of the challenges faced by the significant section of people of SA and SSA as the new-age electronic media presented many opportunities. When we looked at the user-base, Internet growth for the last seven years were the highest in the Asian and African continents, signaling it is yet to be anywhere near maturity (SA and SSA also had the lowest penetrations)³³. The measures of Internet's growing popularity as a media can be based on the growth in the number of people online and for how long, innovations in the applications of Internet, number of new pages of contents, advertisement spent over Internet, etc. And with accessing costs falling and availability rising geographically and over types of devices, the Internet as a rich multimedia offers the best bet to (1) provide education (with supervision and guidance for basic and primary levels) at all levels to (2) provide timely information anywhere, anytime; and more so in SA and SSA.

A study by The United Nations (1997)³⁴ examined possibilities of using the rich, free content of the web in a way, which would benefit the society at large and vulnerable groups in particular. It also took into account the wider framework of actual needs and existing facilities of the Third World 'information-poor' communities. It focused on how Internet was used and how it could potentially be used to build and develop soft-infrastructures like health, education and participation in the political processes. It argued that Internet could only become a tool for social development if it is applied in a way that addressed the complex challenges of improving the lives of the least-privileged and most-needy millions around the world. However, going by all the important questions, Internet has not yet materialized to be a social tool

³³ op.cit. internetworldstats.org. The analysis also showed that Africa and Asia ranked lowest in Internet penetration, at 3.6% and 11.8% respectively. For SA and SSA – excluding South Africa, the penetration would be much lower than these figures as South Africa alone accounted for nearly one-third of African Internet penetration. For South Asia, in its two largest populous nations, India had the maximum Internet users with about 40 million, which is only (3.65 %) of the population whereas Pakistan had about 12 million Internet users with a population penetration of 7.23 %.

³⁴ Uimonen, Paula 'The Internet as a Tool for Social Development' Annual Conference of the Internet Society, INET, 1997 http://www.isoc.org/isoc/whatis/conferences/inet/97/proceedings/G4/G4_1.HTM

of its true potential, though many isolated cases of the success-stories exist. However when the scale is compared with the actual needs SA and SSA, they prove insignificant, at least up till now. In spite of the simultaneous revolution in the user base, its usage and content generations in leading Internet languages, content availability in locally relevant languages of the populace received a low key priority and thereby suffered badly.

Highlighting the language barrier that prohibits Internet content to be relevant for healthcare for the underprivileged, Singh et al (2007)³⁵ in their study on the use of Internet for health and food, estimated that 80%-90% of the health and food-related institutions did not translate their websites into multiple languages, even when the information concerned pandemic diseases such as avian influenza. They argued that although Internet users were often well-educated, there still was a strong preference for searching for health and food information in the local language, rather than in English. It further strengthens our argument of content to be available in local language, even for some of the 'not-poor-but-information-poor' section of people from SA and SSA. That's the only way content can be more effective, and meaningful. In their example, for "avian flu," they found that only 1% of searches in non-English-speaking nations were in English, whereas for "tuberculosis" or "schizophrenia," about 4%-40% of searches in non-English countries employed English. We had similar experience as we found 0.002% content of local places in local languages compared to what's there online for same place in English.

An existing research described four possible interpretations on the definition of digital divide, that potentially lead to information inequality leading to inequality in income in the longer run. For example, the use of telephones and web-enabled computers are measured by the number of devices and the gap on the ability to use ICTs is measured by skill-base and numerous complimentary assets (Fink & Kenny, 2003)³⁶. We strongly feel that populace from two of the least-developed parts of the world, South Asia

³⁵ Singh PM, Wight CA, Sercinoglu O, Wilson DC, Boytsov A, Raizada MN
Language Preferences on Websites and in Google Searches for Human Health and Food Information
J Med Internet Res 2007;9(2):e18
<URL: <http://www.jmir.org/2007/2/e18/>>

³⁶ Fink, Carsten. and Keeny, C J. 'W(h)ither Digital Divide' The journal of policy, regulation and strategy for telecommunications, Volume 5, Number 6, 2003, pp. 15-24

and Sub-Saharan Africa, do face one more critical hurdle compared to most other parts, that of mastering one of the foreign secondary global languages in which rich information is available online at no direct cost. And this medium barrier unfortunately poses a hurdle that majority of underprivileged populace from these two regions isn't prepared to cross even in the medium to longer term; and language deficiency is much deeper than the ability to use ICTs as it's taken for granted in the developed western world.

Table 1: Capability and skill levels required for effective usage of various communication media

	No Literacy	Basic Literacy	High literacy/ Language skills	Computer Literacy	Technical competence
Oral Communication	*				
Radio	*				
Television	*				
Fixed line Telephone	*				
Mobile telephone	*				
Public phone	*				
Newspapers and Printed sources		*			
Fax machine		*			
E-mail			*	*	*
Internet			*	*	*

Source: Pigato, Miria. The World Bank

We, however, believed that the ICT industry, have collectively worked together to operate and interact in the online environment, where finding information has proven to be easier online than in offline platforms. This trend of user-friendliness is observed in all categories of ICT products. We are also of the opinion that high literacy may not be the criteria for effective Internet usage in the days to come. Moreover, there may be a possibility of the local language accessibility of the online content for the vast majority of the populace in SA and SSA in the near future. Existing education level of Internet in these two regions are unlikely to yield insightful findings of Internet usage for future because presently Internet penetration is anyway limited to the well-offs of the society in SA and SSA, who may have anyway completed formal education. Therefore, we don't agree with Pigato when he hinted that higher level skills (language and computer literacy) were only likely to be acquired by those who had completed a formal education, primarily those who have attained literate and numerical skills at a secondary level. We rather feel that even primary level quality schooling can help one finding relevant meaningful content and information

over the Internet provided one is familiar with the language. Language skills, meaning knowing one major Internet language, has increasingly replaced all other capabilities and skill-sets needed to make the most use of online content.

A World Bank³⁷ study (2001) identified the critical access gap between the urban and rural areas, and between the poorest and richest 20% of the population. It found that the new forms of ICTs, including Internet, fax and computers, have touched only some 2% of low-income households, that too mostly in urban areas. The poor therefore mostly rely on the informal networks that they trust, such as family, friends and local leaders for their sources of information, more so in rural areas that are again less competent in using foreign (leading Internet) languages. One way that online content can positively affect the informal networks is making relevant content available in local languages, so that no additional language skills be necessary for local informal communities to reap the potential returns by exploiting the rich educational and information on online content.

Definitions

Going beyond the academic publications, we found the definition of content as *'the sum or range of what has been perceived, discovered, or learned'* (Freedictionary, 2007)³⁸. We liked this definition, and we believed that the impact of learning should be the key aspect in content, if content is to be viewed as a social tool.

Shannon (1948)³⁹ carried out the seminal work when it came to the task of defining information. However, it was more of a mathematical way of looking at information with the measure of information being in bits as continued with the Berkeley study even half-a-century later. His paper laid out the basic elements of digital communication networks with its pioneering focus from its source to its transmission, channel,

³⁷ op.cit. 'Information and Communication Technology, Poverty, and Development in sub-Saharan Africa and South Asia'

³⁸ <http://www.thefreedictionary.com/content> accessed on 29th August, 2007

³⁹ Shannon, Claude E. 'A Mathematical Theory of Communication' *Bell System Technical Journal*, vol. 27, pp. 379-423, 623-656, July, October, 1948,

receiver and finally to the intended destination. For us, we were more interested in the nature of the definition that made the information flow relevant and complete. We took the whole loop backward – starting from the destination, i.e. from the human interface part and moving forward to determine which information out of these ‘sea’ of information would be relevant for the least privileged sections of the destinations in SA and SSA (under-informed sections of population) to help them break the vicious cycle of information poverty.

Barring the human interface part as destination, qualitative nature of information was largely ignored in Shannon’s concept of information. Information as per Shannon was confined to one particular aspect of transmission and storage. Britz (2004)⁴⁰ rather defined ‘Information Poverty’ for situations driven by lack of information *‘as that situation in which individuals and communities, within a given context, do not have the requisite skills, abilities or material means to obtain efficient access to information, interpret it and apply it appropriately’*. From Britz, we don’t get any definition of information; but we understand better the consequences of lack of information, which is relevant for our studies regarding SA and SSA. Loria (2006)⁴¹ too defined information poverty as *‘the absence of vital information necessary for personal or collective development, often due to a lack of information technology or infrastructure, but also caused in many circumstances by an effective censorship of information due to political or cultural factors’*. By ‘Information poverty due to linguistic diversity’, we refer to the soft infrastructural gap which does not allow people to make use of existing online information in other languages (as it’s not available in the language/s they understand). Britz (2004)⁴² defined information poverty as that situation in which individuals and communities, within a given context, do not have the requisite skills, abilities or material means to obtain efficient access to information, interpret it and apply it appropriately. We focus on the interpretation part of it, believing it to be the key area going forward, due to linguistic barriers.

⁴⁰ Britz, J J ‘To Know or not to Know: A Moral Reflection on Information Poverty’ Journal of Information Science, Vol. 30, No. 3, 192-204 (2004) <http://jis.sagepub.com/cgi/content/abstract/30/3/192>

⁴¹ Loria, Par (2006) Religious Information Poverty in Australian Schools Journal of Christian Education, 49 (3), 21-31. http://espace.library.uq.edu.au/eserv.php?pid=UQ:23624&dsID=Loria_UQeSpace_InformationPoverty.pdf

⁴² Britz, Johannes J, ‘To know or not to know a moral reflection on information poverty’ Journal of Information Science, Vol. 30, No. 3, 192-204 (2004)

While talking about poverty and information poverty, we also felt a need to define information poverty in both scientific and economic terms (as poverty is defined in amount of calorie intake/day or on one-dollar-a-day or two-dollars-a-day term). A definition that deals with information intake/day on an annualized/periodic basis by quantifying number of pages read or amount of time exposed to electronic media of information relevance should serve this purpose.

Information utility for economic growth of the underprivileged is more likely when underprivileged communities have timely access to the information in their language and in the format (text, audio, video) they want, just like the value of money is more to the people, who have less of it provided they get it when they need it most (timely credit), in the form of liquidity they desire (cash or otherwise). We increasingly are living in a world, where there isn't any lack of resources (physical, monetary or information/knowledge resources) in the macro-level. In spite of the abundance of physical resources, shortages in soft resources lead to unequal distribution of hard resources in most societies, creating higher and higher degrees of inequality, both in economic and information perspective.

As we proceed to take stock of relevance of the huge online content for the populace of South Asia and Sub-Saharan Africa, we realized the importance of content analysis. Krippendorff (2004)⁴³ viewed content analysis as ways of analyzing meaningful matter such as texts, images, voices (different forms of data), whose physical manifestations are secondary to the meanings that a particular population of people brings to them. Our focus area is the effectiveness on the 'meaning' or 'take away' part that end-users of that content derive, and whether it can provide significant impetus to the population, leading their lives in sub-standard manners to an overall better quality of living. Therefore we realize that medium of communication (or call it information, content) to be of extreme importance for the end-consumer.

It's extremely difficult to categorize (or segment) content in the web for any specific target audience. Any review of the existing content over the web meant for specific countries/regions would therefore be open

⁴³ Krippendorff, DK 'Content Analysis: An Introduction to its methodology' Google Books (2nd edition) http://books.google.com/books?hl=en&lr=&id=q657o3M3C8cC&oi=fnd&pg=PR13&dq=Definition+content&ots=bIgjw4H7xW&sig=QpEMQC6L-gpr5PhMZ9_nvfhOozA

to questions, because web is not bound by any physical boundaries of reach. Without readership patterns and origin/demographic profiles of the readers, it is nearly impossible to meaningfully determine the use of any information over web across the different sections of the people. Growth of the online free content can effectively sustain following ad-based model commercially, where consumers actually get sponsored by the advertisers to view content, free of cost in the Internet. The growth of the Internet and its increased acceptance amongst the content generators, content consumers and advertisers so far has validated and strengthened this relationship that effectively provides end-consumer with endless content at insignificant bare minimum direct or indirect costs. One can reasonably be hopeful that the same model, valid for major Internet languages, would be sustainable when it comes to local language online content as well, applicable for diverse languages of SA and SSA origin. It effectively is a chicken-and-egg story where readership would emerge provided local language content is good, and content availability (thro' translation or generation) gets impetus from readerships.

Research Methodology

We applied two different methods for these two backward regions of the world. For SSA, we checked whether linguistic diversity poses an additional hurdle for the nations with highest linguistic diversity within SSA, in meaningfully educating their populace, which in the long run would result in higher per capita income. Part of our inputs came from couple of existing studies of ten most linguistically diverse nations of SSA⁴⁴ that accounted for nearly 25% of SSA population (Table 3 without the last four columns, which we added). We surprisingly found that in eight of the ten cases, these nations lagged the average GNI (both absolute and PPP, in one case PPP GNI was not available) for Sub-Saharan Africa. Table 3 provides the findings qualitatively.

For South Asia, an interesting development and product launch by Google in August 2007 helped us search online content in few of the local languages and compare the quantum of local language content with the 'sea' of the English language content. We realized that the number of search results (web pages carrying the keyword that we searched for) would be a very crude measure to judge quantitatively and

⁴⁴ op.cit. 'Ten most linguistically diverse countries in sub-Saharan Africa'

qualitatively the content in different languages, but as stated earlier; no research can be accurate in this dynamic huge area.

The objective was to check the linguistic diversity and its effect in the backward regions of the world (for SSA) and on online content availability in the local languages (for SA). The limitations of the available resources and linguistic barriers didn't allow us to carry both the studies for both these two regions separately. However, keeping in mind the broad commonalities in terms of economic parameters and education levels of societies in both SA and SSA; we felt that we can reasonably take both these two different individual findings to be true for the other regions as well.

First, we looked at the linguistic diversity of SA and SSA. India, the largest nation in South Asia and accounting for nearly 70% of the population in SA is a unique case as there were 22-officially recognized languages as per Indian constitution. Moreover, there are official languages at the state and center level, and there's been no national language. The 1991 census-data of Indian government recognized 1,576 classified "mother tongues", whereas Ethnologue listed 415 living "Languages of India" (out of 6,912 worldwide). And as per the same 1991-census data, 22 languages had more than a million native speakers, 50 Indian languages had more than 100,000 and 114 had more than 10,000 native speakers.

Pakistan, the 2nd biggest nation in South Asia again had 4 provisional languages, 2 regional languages, 2 official languages and one national language (Urdu, though only 8% of Pakistanis speak it as a 1st language against 44% for Punjabi), and another 80-speaking tongues⁴⁵. Thankfully the diversity in languages is less in the other five South Asian nations although Sri Lanka also has diversity within two of its largest communities. Image 1 provides this linguistic diversity of South Asia (we counted minimum 26 languages here with some understanding of Indian languages).

⁴⁵ Wikipedia, languages of Pakistan http://en.wikipedia.org/wiki/Languages_of_Pakistan accessed on 29th August 2007

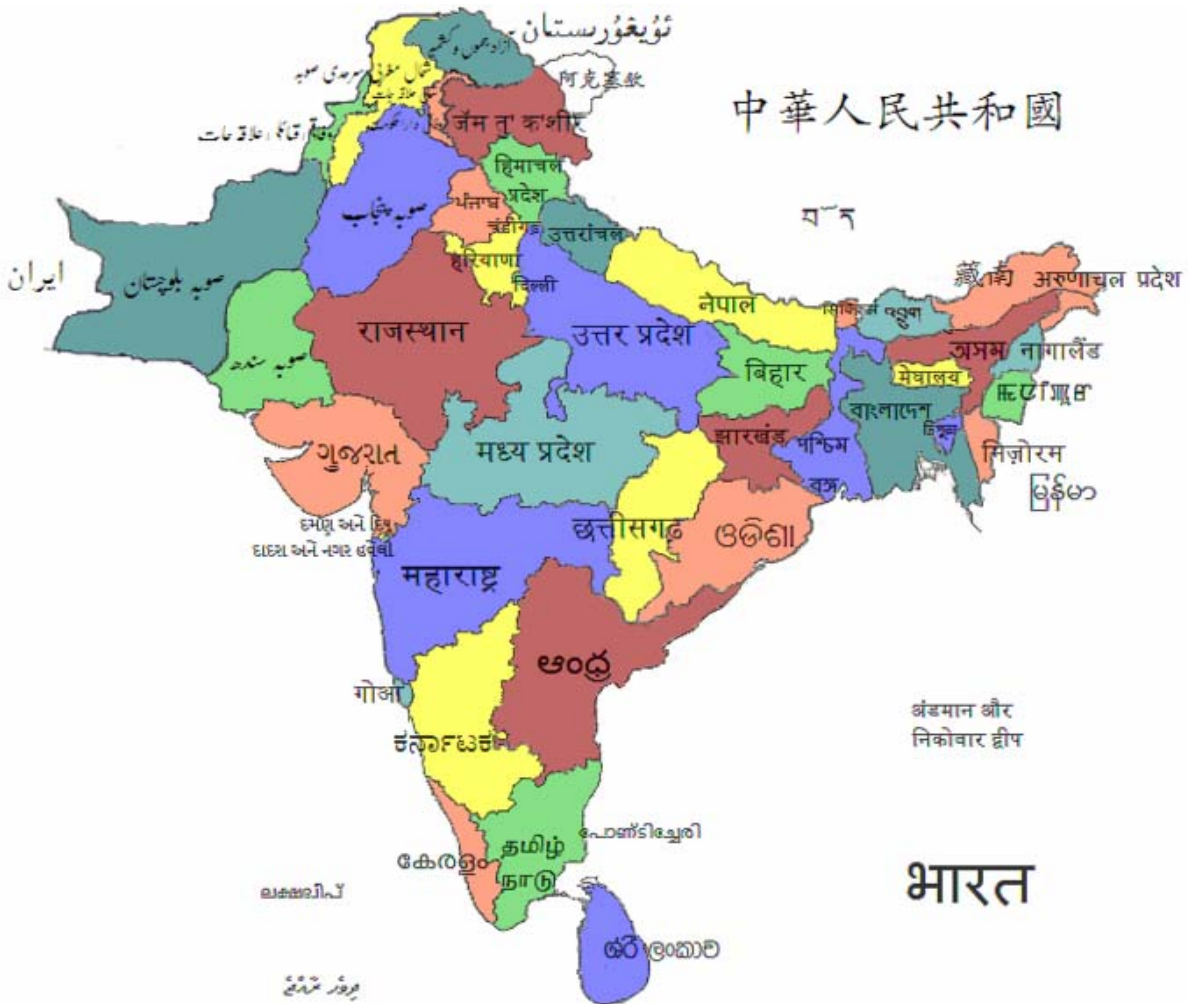


Image 1: South Asia local languages (Source: Wikipedia, http://en.wikipedia.org/wiki/Image:South_asia_local_lang.PNG)

With the ‘global village’ concept increasingly becoming a reality with globalization; it’s imperative that the members of the ‘global village’ have uniform and equal access to the flow of the global and local information. As the world increasingly adopts only a few major languages for communication; there is a real danger that a large section of the people in the ‘global village’ stays away from meaningful economic activities as they can’t comprehend the business language that this global village speaks. An increasing number of languages are getting endangered with the process of globalization⁴⁶; and knowledge creation through continuous content generation in many other local languages beyond these few leading global languages, suffer, challenging their survival in future. We found it to be true with our experiments with

⁴⁶ op.cit. ‘The Atlas of the world’s languages in danger of disappearing’

couple of South-Asian languages.

Linguistic diversity as observed in SA is equally visible in SSA. Swahili, the most widely spoken language in SSA is used by 5-10 million people as their native language; although it's used by about 80 million speakers (nearly 10% of the population of SSA)⁴⁷. French, English and Arabic are also used along with local languages as official medium in many nations⁴⁸. For Sub-Saharan Africa, we applied a different methodology in gauging whether linguistic diversity has any direct bearing on the per-capita level GNI of nations and also on the Internet penetrations. For that we adopted ten most linguistically diverse countries in Sub-Saharan Africa in 2000-2001, and latest Internet penetration levels to examine whether linguistic diversity within low-income nations ($\$650 < \text{per-capita GNI} < 3051$, absolute figures) of SSA posed further challenges by making these ten nations even more backward. There is a scope to apply statistical hypothesis to check whether linguistic diversity indeed causes lower per-capita GNI in SSA context (by examining the means of sample and population, however we also observe that this hypothesis does not hold strong within the sample itself), however at this stage we have refrained from applying any quantitative tools in our predominantly qualitative studies.

Findings for South Asia

We conducted our experiment of local language content in South Asia in two of leading South Asian languages – in Bengali and in Hindi. A basic understanding of these two languages from their native speakers' perspective is therefore important. In Bangladesh, 98% of the population had 1st language as Bengali⁴⁹ and in West Bengal, a province of India, 85%⁵⁰ of the population were Bengali native speakers, whereas globally, it is ranked at number 4th after Chinese, English and Spanish in terms of top languages⁵¹. The similar figure (native speakers) for Hindi due to the regional diversity in India is much

⁴⁷ http://en.wikipedia.org/wiki/Swahili_language

⁴⁸ 'Is number of speakers really so important' <http://www.antimoon.com/forum/t8181-0.htm> accessed on 3rd September, 2007

⁴⁹ Bangladesh _ Language, Culture, Customs, and Etiquette <http://www.kwintessential.co.uk/resources/global-etiquette/bangladesh.html> accessed on 24th August 2007

⁵⁰ Bengali Orientation <http://www.everyculture.com/South-Asia/Bengali-Orientation.html> accessed on 24th August 2007

⁵¹ Most widely Spoken Languages, Summer Institute for Linguistics (SIL) Ethnologue Survey (1999)

less (20% of Indian population) as only 180 million (1991 figure) of Indians did have their native language as Hindi⁵². However, Hindi is bestowed with the official language status in India with its nearly 1.1 billion people.

The event of Google launching its local language supports in 14 Indian languages did trigger this experiment. We picked two of the largest cities of Bangladesh, two of the largest cities of West Bengal (both having Bengali as predominant native language) and the capital of India as locally relevant keywords. The search with 'কলকাতা' (needs Bengali script support for viewing, 'Kolkata' in Bengali) resulted 933 matches, whereas the English keyword ('Kolkata') resulted in 9,730,000 matches in spite of census figure showing a predominant overdependence of Bengali as the language for communication in and around West Bengal. A repeat of this exercise with 'ঢাকা', (Dhaka) the capital of Bangladesh resulted in 46,300 matches whereas 'Dhaka' in English resulted in 6,550,000 matches. The local language contents mostly originated from (1) Wiki-forums, (2) Government sites and (3) Media sites. And accepting all criticism, we believe that the ratio of Kolkata to Dhaka on how they fare in a Google search in local Bengali language vis-à-vis English tells us few insightful things:

- Dhaka, being the national capital and national language of Bangladesh being Bengali with 98% usage amongst population possibly explains significantly more web-content for Dhaka in Bengali than Kolkata in Bengali. When we examine content in Bengali using Google search for Bengali and English; we see that (local language (Bengali)/English content) for Dhaka to be 70 times more than same for Kolkata; though 'Kolkata' has more content (in terms of web-pages) in English. This clearly highlights less local language content for Kolkata in Bengali due to linguistic diversity that India faces compared to Bangladesh, where 98% people speak same language.

We thought of carrying a similar search with the commercial capital of India, namely Mumbai. However as Mumbai changed its name from Bombay, we needed to carry it out in four keywords; and therefore restrained from doing it (any minor spelling changes in Hindi during keyword searches would throw

<http://www2.ignatius.edu/faculty/turner/languages.htm> accessed on 24th August 2007.

⁵² Ethnologue report for language code: hin http://www.ethnologue.com/show_language.asp?code=hin

different number of web-pages). Table 2 summarizes these findings of content in local languages against English-language content:

Table 2: Keyword search results in English vis-à-vis local languages (Search engine used: Google)⁵³

Keyword search	Search result in English	Search result in local language	Ratio of web-content for populace* in local language to English
India	323,000,000	1,120,000 (in Hindi 'Bharat')#	0.003
Bangladesh	93,800,000	79,000 (in Bengali)	0.0008
New Delhi	45,600,600	184,000 (in Hindi, 'Naya Dilli')	0.004
Dhaka	6,550,000	46,300 (in Bengali)	0.007
Kolkata	9,730,000	933 (in Bengali) & 47,900 (in Hindi)	0.00009 (for Bengali to English) & 0.005 (Hindi to English)
Chittagong	1,570,000	21,000 (চট্টগ্রাম in Bengali)	0.013
Siliguri	893,000	15 (in Bengali)	0.00002

* This may be a crude measure of content; however this is based on number of web-pages with keyword search results.

As India is a country of diversity, true for languages as well; a better measure would be to search 'India' in all leading Indian languages. Cities from Bangladesh clearly scores better in Bengali content compared to cities from West Bengal, whereas Bharat ranks better than Bangladesh. This could be due to the fact how we spelled Bangladesh in Bengali.

Findings for Sub-Saharan Africa

Here we collated our information from three different secondary sources- linguistic diversity, GNI and Internet Penetrations. Essentially these three sources captured (1) Top ten nations amongst all nations in SSA in terms of Intensity of linguistic diversity and their measures, (2) GNI levels and (3) Internet penetration levels. We acknowledge that the 1st data source belonged to the year of 2001, whereas the other two were of 2006 and 2007. We assumed that the distribution of native language speakers in SSA in 2006 remained similar to 2001. The findings are shown in Table 3 below.

⁵³ Rao, M 'Struggling with the Digital Divide: Internet Infrastructure, Content, and Culture' Ontheinternet, Isoc.org 2000 <https://www.isoc.org/oti/printversions/1000rao.html> suggested amount of subnational content (about states, provinces and cities; and we followed cities in local language) as one of the seven measures to judge maturity of Internet content in a country.

Table 3: GNI and Internet penetration level for ten most linguistically diverse SSA nations

Country	Pop. mill. 2000	No. of languages	Official langs.	3 largest langs. (year)	No. mother tongue speakers	Larg-est lang. grps. as % of pop	Mean of 3 largest lang grps as a % of total pop	No. lgs. per million inhabitants (pro rata)	Per Capita GNI (\$)	Per Capita GNI PPP (\$)	Rank PPP	Internet penetration (% of population)
Cameroon	14.9	282	French English	Fulfulde, Adamawa (1986) Ewondo (1982) Yemba (?)	668,700 577,700 300,000	4.5 3.9 2.0	3.5	18.9	1080	2370	165	1.4
Central African Republic	3.7	69	Sango French	Sango (1988) Manza (1996) Gbaya, NW (1996)	350,000 220,000 200,000	9.5 5.9 5.4	6.9	18.6	360	1280	186	0.3
Chad	7.9	132	Arabic French	Arabic, Chadian Spoken (1993) Ngambay (1999) Kanembu (1993)	754,590 750,000 389,028	9.6 9.5 4.9	8.0	16.7	480	1230	188	0.4
Tanzania, United Republic of	35.1	135	Swahili	Sukuma (1993) Gogo (1992) Haya (1991)	5,000,000 1,300,000 1,200,000	14.2 3.7 3.4	7.1	3.8	350	740	205	1.0
Uganda	23.3	43	English	Ganda (1991) Nyankore (1991) Chiga (1991)	3,015,980 1,643,193 1,391,442	12.9 7.1 6.0	8.7	1.8	300	1490	181	1.7
Congo, Democratic Republic of	50.9	218	Kongo Lingala Luba-Kasai Congo Swahili French	Luba-Kasai (1991) Mongo (grp of lgs) (1993) Luba-Katanga (1991)	6,300,000 4,800,000 1,505,000	12.4 9.4 3.0	8.3	4.3	130	720	207	0.2
Mozambique	18.3	39	Portuguese	Makhuwa (1996) Tsonga (1989) Lomwe (1991)	2,500,000 1,500,000 1,300,000	13.7 8.2 7.1	9.7	2.1	340	1220	189	0.7
Côte d'Ivoire	16.0	77	French	Baoulé (1993) Senoufo, Cebaara (1993) Dan (1993)	2,130,000 862,000 800,000	13.3 5.4 5.0	7.9	4.8	870	1550	179	0.6
Liberia	2.9	29	English	Kpelle, Liberia (1991) Bassa (1991) Mano (1995)	487,400 347,600 185,000	16.8 12.0 6.4	11.7	10.0	140	NA	NA	0.03
Benin	6.3	51	French	Fon-Gbe (1993) Yoruba (1993) Bariba (1995)	1,400,000 465,000 460,000	22.2 7.4 7.3	12.3	8.1	540	1160	191	5.5
SSA Average									842	2031		3.6%*

Source: population figures - UNPD 2001; language information & GDI - Ethnologue 2000 CDROM, (<http://sociolingo.wordpress.com/2007/04/18/ten-most-linguistically-diverse-countries-in-sub-saharan-africa/>) and GNI from <http://siteresources.worldbank.org/DATASTATISTICS/Resources/GNIPC.pdf> (2006 figure), and Internet Penetration from Internet World Stats - 2007 (<http://www.internetworldstats.com/stats1.htm#africa>) accessed on 31st August. * figure is of all of Africa, however as SSA accounts for nearly 85% of Africa's population, we took it as an indicative measure South Africa alone however accounted for almost one-third of all African Internet connection in 2000 ('For Most Africans, Internet Access Is Little More Than a Pipe Dream, <http://www.ojr.org/ojr/workplace/1079109268.php>)

Analysis of our findings in light of prior research

The indicator of Information inequality could be many. In spite of the awesome growth in the Internet content, an audit by Wendy & Francisco (2000)⁵⁴ showed that many underserved communities (in the U.S.) were not benefiting fully from content because of the barriers faced related to online content. He categorized these barriers as (1) Scarcity of information for the development of local community, (2) literacy limitations (3) language barriers & (4) the lack of cultural diversity in Internet content. Though we see interdependency in these four barriers, both literacy limitations and language barriers are their worst-off for SA and SSA (infrastructural bottlenecks being the hard side of it).

The demographic analysis of people of both South Asia and Sub-Saharan Africa showed that a significant majority don't use colonial language legacy in their communications. Population and language related studies have also been equally difficult (like number of web pages language-wise), and estimation varied widely. Although Yaleglobal (2004)⁵⁵, citing The Guardian (and a survey conducted by India Today) proclaimed India to be the home of world's largest English speaking population; we doubt that figure, when it comes to making sense of English textual content by Indian populace keeping in mind the abysmal level of literacy in India and its quality of basic education. We would rather go with the figure given by Wikipedia⁵⁶ (100 million i.e. less than 10% of the population including 2nd language and 3rd language although 1st language speakers were hardly 0.2% of that number). Qualitatively it won't be wrong to say similar situations prevailed in other nations of SA and SSA.

As we examined the information overflow in terms of online content for one section (the online populace that communicates in English or in any other few major Internet languages), and almost an absolute lack of information for the other significant non-English speaking sections which may be connected, (if not now

⁵⁴ Lazarus, Wendy; Mora, Francisco. 'Online Content for Low-Income and Underserved Americans: The Digital Divide's New Frontier. A Strategic Audit of Activities and Opportunities' 2000, Educational Resources Information Center (ERIC), US Department of Education. http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/16/2a/3d.pdf

⁵⁵ Yaleglobal (2004) 'Subcontinent raises its voice', it estimated a 350 million people within India to speak English

⁵⁶ http://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population accessed on 3rd September. It also had figures for other nations from SA and SSA.

in near future, from South Asian and Sub-Saharan African population, not knowing any major Internet languages), we came across one initiative of globalvoicesonline.org, a project initiated a *lingua-project*⁵⁷ to translate existing contents in six leading languages (Bengali, Spanish, French, Portuguese, Chinese (simplified) and Chinese (Traditional)). The *lingua project* stated in its site that it had plans to initiate local content developments in three more languages (Farsi, Russian, and German). However, barring Bengali and Traditional Chinese; all the four other existing languages already ranked in the top ten Internet languages; and so do Russian and German. Though we were encouraged by this project and believe more such projects to be the need of the hour (with more resources), we also were disappointed by the quantum of local language content development in Bengali (too less, almost insignificant) and also by lack of any other South Asian or Sub-Saharan languages. It focused primarily in languages where existing online content as adequate, the only exception being Bengali.

What came as a sign of hope are initiatives like local language blog supports as initiated by Google (e.g. blogs in Hindi on Google platforms)⁵⁸. With more and more languages getting added, this can be a big boost to the problem; however acceptance of bloggers from these regions to write in local language does face other challenges (practice, keyboard familiarity and readership issues).

Policy Recommendations and Conclusions

We identified linguistic diversity to be an important characteristic of socio-economic backwardness in SA and SSA, as it creates hurdles to information flow uniformly. We thereby believe that adequate resources in proportion to other drivers (focusing on the physical parts of connectivity) universal access is to be deployed for content development in local languages – be through automated software based technologies or through new initiatives that encourages original local language content developments (like globalvoicesonline.org), more so when no global content in other languages serves local purpose.

⁵⁷ <http://bn.globalvoicesonline.org/lingua-project/>

⁵⁸ Official Google Blog ‘Now you can blog in Hindi’ at <http://googleblog.blogspot.com/2007/04/now-you-can-blog-in-hindi.html> dated 12th April, 2007. To state here, one of the authors of this paper who blogs with Google platform saw that titles of one of his blogs posted in November’07 was translated in Hindi though the author kept it in English, and didn’t opt for any translation/change in medium. Also ‘Google Adds English to Hindi Transliterator in Blogger Editor’ at http://labnol.blogspot.com/2007/03/google-adds-english-to-hindi_5338.html dated 8th March 2007 explains many of these developments.

And unless this gets done; SA and SSA would miss 'content for development' and thereby will fail to integrate with the global economy as others take more advantages from this online content revolutions and surges ahead.

An increasing amount of debate has lately focused not only on the economic growth that the world can sustain, but more on the quality and 'inclusiveness' of that economic growth for the vast number of people, still living under poverty. The picture gets even more skewed when we examine growth in content and 'inclusiveness' of that content from its point of utility for human development. We simultaneously observe information abundance for the well-informed in a rising 'sea' of information side-by-side with information poverty for the uninformed or ill-informed people handicapped by their inabilities of not knowing any major Internet languages, which again number in billions all over the world. However, what largely remains ignored is the attention this language-barrier to effective utilization of online content gets from policy-makers, to content generators, to other service providers. Linguistic diversity, rich in terms of cultural heritage, can also be a problem for the communities, more so when the community is characterized by poor education, literacy levels and extremely poor access to information. On the other hand, online content has witnessed phenomenal growth driven by new ad-based revenue models. The challenge, therefore, is how we can identify, convert and archive (online) part of the global online content-pool relevant for underprivileged and under informed sections of people in SA and SSA in their local languages, supported by search engine facilities again in local languages. The task may look daunting; however with the ongoing developments, innovations and a little policy-focus; this can be made a win-win proposal for all stake-holders in the value chain, and thereby be made commercially viable as well. The alternative of educating populace from SA and SSA with another language skill would prove to be much more difficult.

Our findings faced limitations as we knew only couple of the major languages of South Asia (and none of the languages of SSA). Therefore we could not gauge local language content in major languages of SSA using same method (provided Google supported that initiative as it did for Indic-languages). There have also been diverse developments starting with software that can automatically convert content from one

major Internet languages to another. Though we looked at many of these latest developments, it's nearly impossible to be certain that we covered all dimensions of content conversions in this paper.

Before we end, we believe that in spite of various ongoing developments and innovations; the problem should not be left unattended at the whims of market forces. We recommend a dual approach where policy-makers identify all members in the online content value-chain (starting from content generators to end-consumers), work with them closely to identify (through a degree of artificial intelligence or automation) contents relevant for SA and SSA, translate those in all relevant local languages, store the translated versions, and provide local language search facilities to integrate with the end-user. In certain cases, policy-makers would have to encourage creation of local language content where there's a gap on suitability of global ones with local requirements. With little impetus, the model can be self-sustaining in these local languages as well as we have seen with online content itself.