Making the 'Next Billion' Demand Access*

The Local-Content Effect of `google.co.za` in Setswana

Bastiaan Quast†

Abstract

This paper shows that an exogenous increase in accessibility of local language content leads to an increase in demand for Internet connectivity among native speakers. Internet connectivity provides enormous improvements in quality of life as well as opportunities for the newly connected, yet recent attempts to connect the current 'next billion' in places such as sub-Saharan Africa have not met expectations. In places where infrastructure has come online and prices have gone down, the expected consequent increase in usage was not observed. The introduction of the Setswana language in the South-African Google Search website was a spillover of the Botswana Google search website being translated from English to Setswana. This exogenous improvement in the accessibility of Setswana-language content has resulted in a substantial increase in the number of native Setswana speakers coming online and owning personal computers. It has also led to increased usage of the Setswana language online, creating a positive-feedback loop. This suggests that connecting the fourth billion will require a greater focus on the demand-side of connectivity, specifically by means of local content.

1 Introduction

The term 'Connecting the Next Billion' was introduced in The Economist's 2006 'End of Year Report' (Standage, 2006), discussing the infrastructural requirements for doing so. Since then, close to 2 billion people are estimated to have been connected to the Internet, up from the just over one billion at the time of writing (Sanou, 2015). However, it is unlikely that the current 'Next Billion' will be connected as easily as the previous ones.

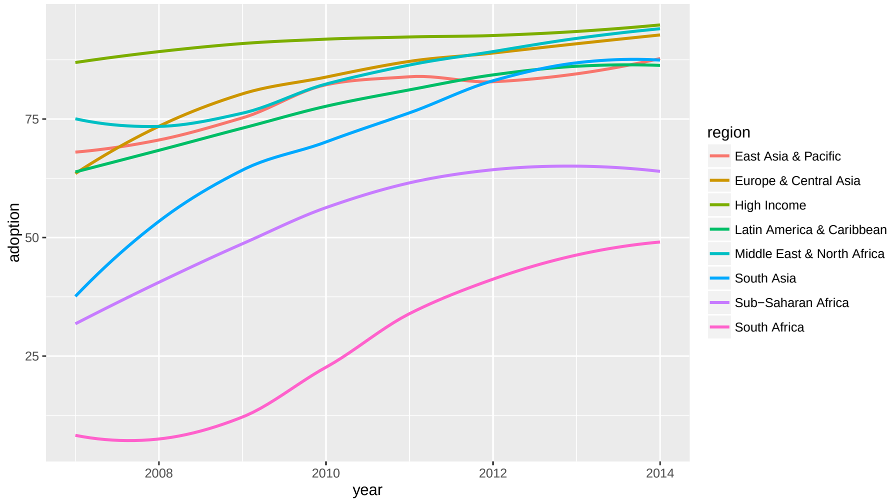
In the period 2010-2014 the average annual growth of Internet bandwidth in sub-Saharan was over 50%. This increased bandwidth also caused a corresponding drop in the price of Internet access, bringing the sub-Saharan average of a 500MB prepaid Internet bundle down to around $10, thereby putting it within range of the emerging middle classes. Yet, despite increased range and improved affordability, sub-Saharan Africa is showing stagnation in the growth of Internet connections.

*http://qua.st/internet-access

†http://qua.st | bquast@gmail.com | bastiaan.quast@graduateinstitute.ch | Maison de la paix, Geneva, Switzerland

As is shown in Figure 1, mobile Internet growth, the oft mentioned key driver of Internet accessibility in sub-Saharan Africa, is showing stagnation to the point of being near level or even slightly negative. Unlike in other regions in the figure, this observed stagnation is not a consequence of market saturation, as adoption levels in 2015 are only just over 60 percent.

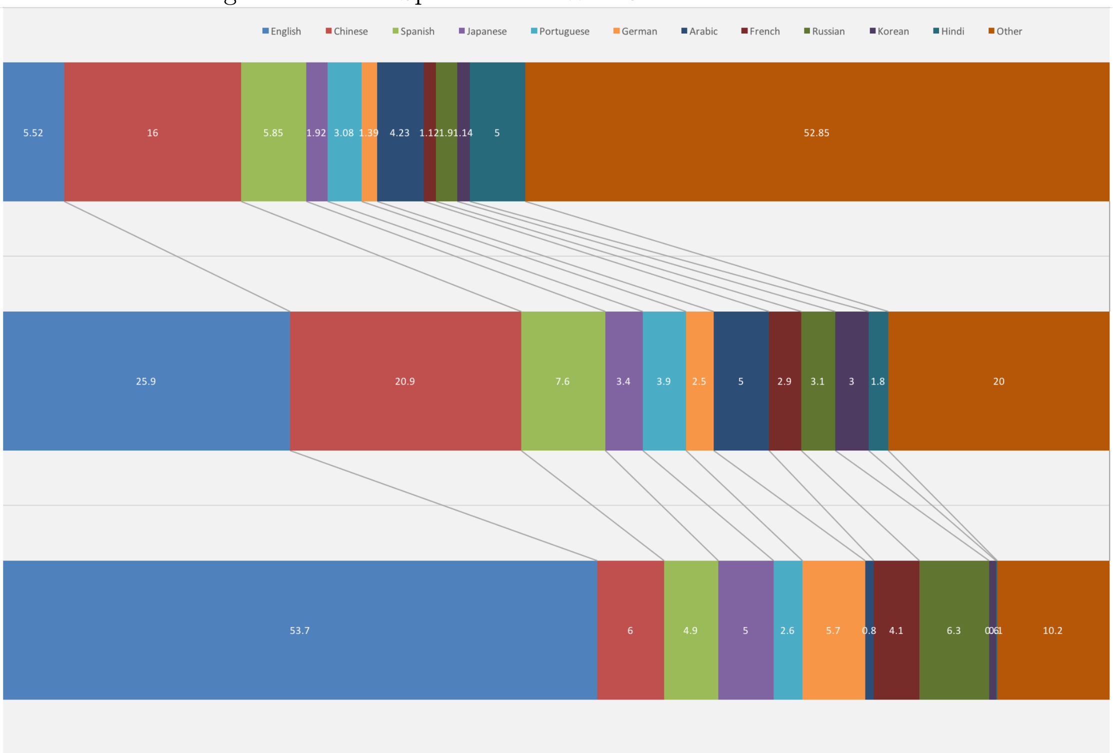
Figure 1: Mobile Internet Growth



The dynamics of internet adoption and content creation take the form of a two-sided market, wherein the two sides benefit from cross-side network effects. In this case the sides are content creators, such as news websites, and content consumers, the users. Thus adoption should follow a virtuous circle, whereby the content offering encourages more users to come online, which in turn incentivices more content creation, unfortunately this virtuous circle does not always start for every language. Herein also lies the difficulty with finding empirical evidence of these dynamics, since the process of adoption by users and content creation is inherently endogenous. The paper seeks to empirically address the question if increased access to content does indeed lead to an increase in Internet adoption.

The increased range (coverage is 99.79%) and affordability, taken together with stagnant growth, suggest that connectivity is lagging because of lower-than-expected demand, as a result of insufficient perceived value. Part of this inadequate value proposition is that as Internet usage becomes more widespread, potential Internet users have lower than before affinity to the 'global languages' that are currently dominant on the Internet. The amount of Internet content in certain languages is very disproportionate to the number of native speakers that are online. The current offering of online content, primarily in a small number of languages, therefore provides insufficient incentive for many to seek access to the Internet. In Figure 2 we can see that while on the one hand, English represents a relatively small percentage of the global population's native speakers (around 5%), it is already a much larger number of Internet-connected global population (around 25%), but at the same time represents more than 50% of all the online content. On the other hand, Hindi is also the native language for about 5% of the global population. However, of the online global population, it is the native language for about 2%, yet the number of websites in the Hindi language is under 0.1% of all web content. The ‘other’ languages in this figure represent over half the global population, as can be seen in the top bar and around 20% of the Internet-connected population, yet all of these languages combined only account for around 10% of online content.

Figure 2: Native Speakers and Web Content



Much of the research on Internet and language focusses on the preservation of smaller languages, in particular indigenous languages (Gandal, 2006). Viard and Economides (2014) use macro level connectivity data and a model whereby countries that share languages are used to isolate the effect of content on demand for connectivity, which is positive and significant. Previously the ITU also highlighted the importance of language availability to improving demand for connectivity (PeÃÂ±a-LÃÂ^3*pez*, 1999).

Sinai and Waldfogel (2004) show that expanded Internet usage in cities can lead to less racism, an issue that is particularly relevant to South Africa. Jensen and Oster (2009) finds that the introduction of cable television in Indian states has a pronounced positive effect on attitudes towards the oppression of women, violence against women, son preference and as well as decreased fertility, as Internet provides an in some ways similar window on the outside world, we might expect some of these things to also follow widespread Internet adoption.

Internet connectivity is a two-sided market, user demand for connectivity is a function of relevant content availability, conversely, content creation and availability is dependent on the amount of interested users. As such, the formation of both vicious and vituous circles between these two is possible. Furthermore, this feedback makes it difficult to empirically isolate a causal effect of content availability on demand for Internet connectivity.

This paper exploits an exogenous increase in the accessibility of local content in the Setswana language in South Africa, in order to isolate the increase of Internet usage among native speakers. In 2010 Google collaborated with a Botswanan team of linguists (Otlogetswe, 2010) to make its Botswanan website (`google.co.bw`) available in the local language: 'Setswana'. In addition to being spoken in Botswana, there is also a sizable population of Setswana speakers directly across the border in South Africa, where it is also one of the official state languages. This led to the Setswana-language interface also being introduced on the South-African Google website (`google.co.za`), as a spillover of the translation work originally performed for Google's Botswanan website. This introduction led to a large increase in the number of native Setswana speakers reporting to have spent some amount of money in the past 30 days on Internet access.

The Google Search interface represents a very small number of words on the Internet and it is not required to use a certain interface language in order to search for content in this language. Yet, the search page is in many cases the first website viewed by users and thereby has a substantial impact on the decision to further engage or not. Asides from being able to understand the interface of the website, having this interface be in a local language also encourages usage of this local language, which in turn reveals more local language content.

In short, we can identify two main channels through which this promotes increased online engagement, which together constitute the theory of change. Firstly, the ability to read and understand the words of the interface increases the chance that a user continues using the website and the Internet at large. Secondly, the visibility of local language content increases the likelihood of the user entering search queries in the local language and thereby finding more content in the local language.

The vast majority of Internet access in developing countries is through handheld devices such as smartphones. However, due to the limited 'real estate' (surface area of the screen) on a mobile website, the link to changing the interface language is replaced with a dropdown menu that reveals the additional language options (see screenshots in Figure 10). Generally, the website will default to the operating system (Android / iOS) language, however, since many local African languages are not available as a system language, this is not possible there, leaving the website to revert to its default (in South Africa: English). The fact that the introduction also led to increased desktop-computer ownership but not mobile-phone ownership further substantiates this hypothesis.

The data used for this study comes from the South African National Income Dynamics Survey, provided by Southern Africa Labour and Development Research Unit (2008, 2012, 2013), the data is further discussed in section 2. After which section 3 discusses the methods employed in this study, specifically, the discussion of the identification strategy can be found is subsection 3.1 and the use of the Difference-in-Differences estimator in subsection 3.2. Further, we present the results of the estimation in section 4. Finally, we conclude in section 5.

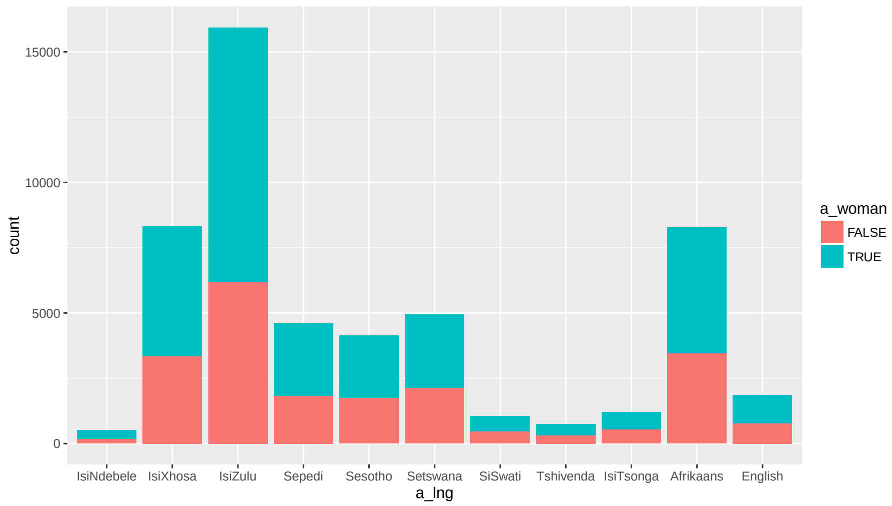
2 Data

South Africa's National Income Dynamics Survey collects data on a representative set of around ten thousand households over time. The first survey took place in 2008, the second one in late 2010 and early 2011, and the third one took place in 2012 (Southern Africa Labour and Development Research Unit, 2008, 2012, 2013). It contains an extensive household questionnaire, which breaks household expenditure down into many forms of food and non-food expenditure. In addition to this, the household income is calculated and imputed with other income such as home ownership. The individual (adult) questionnaires also contain information on linguistic skill in English and in the native language, as well as a series of variables relating to communication technology ownership and utilisation. In Figure 3 we can see the number of native speakers of each in the dataset, coloured by the sex of the individual. The dataset contains 51,612 observations (adult individuals), of which 2,806 are female native Setswana speakers and 2,140 are male native Setswana speakers.

A detailed breakdown of the mean number of households that spend on Internet access and cellphone usage, as well as cellphone and computer ownership by language and time period (wave) is presented in Table 4 of section 7.

Figure 3: Native Language and Sex

```
ggplot(data=NIDS, aes(x=a_lng, fill=a_woman)) + geom_bar()
```



3 Methods

This section begins with a discussion of the identification strategy employed, followed by a explanation of the estimator used to operationalise this, and concludes with a description of the software used for this estimation.

3.1 Identification Strategy

This paper exploits the introduction of the Setswana interface language to Google Search in South Africa as a spillover of the development of that interface for the Botswanan Google Search website. By comparing the number of native Setswana speakers in South Africa being Internet users, with the number of South Africans with a different native language around the same time, we isolate the effect of this introduction.

The Setswana language was first developed for the Botswanan Google Search website (`google.co.bw`). As such, the introduction of Setswana to the South African Google Search (`google.co.za`) was a spillover effect of that development. This allows us to rule out any possible endogeneity issues that might otherwise arise in contexts such as these. For instance, the Afrikaans language is almost solely spoken in South Africa. When we observe that the introduction of the Afrikaans Google Search interface occurs around the same time as a growth in the number of native Afrikaans Internet users, it will be hard to isolate the effect from the introduction from its cause (since an increase in native Afrikaans Internet users would be a good reason to introduce it as an interface language).

Substantial numbers of Setswana speakers exist in Botswana, South Africa, Zimbabwe, and to some extend Namibia. However, the language is most important in Botswana, where it is spoken by approximately 80% of all people, and where it is the only official language other than English. As such, it is also the place where most linguistic work on the Setswana language takes place. The Setswana Google Search interface was also developed at the University of Botswana by prof. Otlogetswe.

It is worth noting that it is very common not to personally own a computer and 'paying for Internet access' therefore also includes a lot of people who use the Internet in other locations such as Internet cafe's.

In addition to using the propensity to spend on Internet (in the last thirty days), we also use the propensity to own a computer as a dependent variable.

3.2 Regression Specification

As mentioned in the above section, we compare the change in the level of Internet users among native Setswana speakers in South Africa, with that of native speakers of other language in South Africa around the introduction of the Setswana interface to the South-African Google Search. For this we use a Difference-in-Differences estimator (Abadie, 2005; Imbens and Wooldridge, 2009) using a native-Setswana speaker dummy variable (`setswana_logical`), interacted with an event dummy variable (`interface_intro`). The former is `TRUE` when the native language of the individual (`a_lng`) is `Setswana` and `FALSE` otherwise. The latter is `FALSE` for data collected prior to the introduction of the Setswana interface language (late 2010, here wave 1 and 2) and `TRUE` after this introduction (here wave 3). The model than takes the form as described in equation (1).

$$y_{it} = \alpha_i + \lambda_t + \delta D_{it} + \beta X_{it} + \epsilon_{it} \tag{1}$$

Where α_i represents the individual fixed effects, λ_t represent the time fixed effects, and X_{it} are the time varying covariates. The ϵ_{it} is the error term.

Finally the term of interest is D_{it} which represents the treatment effect.

In addition to this estimation, we use an alternative specification whereby a factor variable of native language is interacted with the event dummy variable. In a linear regression context, factor variables are estimated as dummy variables for all levels (here: all languages) except for one 'base' level, which is where all language dummies are `FALSE` (i.e. `0`) and the level (native language) thus has to be the first one (here `IsiNdebele`).

The `h_nfnet` variable is recorded at a household level, as such a standard error correction needs to be applied, the model with standard error corrections is reported in the appendix.

Lastly, the dependent variables are both logical or binary variables, as such, normally a model such as logit should be used. However, since I am using Difference-in-Differences, this model would be undefined.

3.3 Software

The estimation is primarily performed using R (R Core Team, 2016), specifically using `lm()` and `glm()` functions included in the `stats` package (Venables and Ripley, 2013). Additionally, I make the of the `plm()` and `pglm()` functions which are available in packages by the same names (Croissant, 2013; Croissant and Millo, 2008). Standard error corrections are computed using the `lmtest` and `sandwich` packages Zeileis (2004, 2006); Zeileis and Hothorn (2002).

In order to make the result as easily reproducible as possible, this research and writing in the article has been done exclusively using open-source software such as R (R Core Team, 2016). This document is written and LyX (LyX Team, 2016) in the LaTeX(Lamport, 1985) language and compiled using the LuaTeX implementation(Hoekwater et al., 2016). The integration of R code and output in the document is performed using a process call literate programming Knuth (1984) using the knitr implementation Xie (2015) of the Sweave framework (Leisch, 2002).

All changes are logged using the version control system Git (Git Team, 2016) and publicly available on GitHub at https://github.com/bquast/Making-Next-Billion-Demand-Access/[1]

4 Results

In the base model, we use an interaction of the `interface_intro` dummy and `setswana_logical` dummy in order to isolate the effect on the explanandum, a dummy variable describing household expenditure on Internet in the last thirty days or not (`h_nfnet`, household non-food Internet). The results of this estimation are presented in Table 1.

We find that the interaction term of the event dummy (`interface_intro`) and the native Setswana speaker dummy (`setswana_logical`) is positive and highly significant, with a p-value around `0.0018`. Both the individual dummy variables (`interface_introTRUE` and setswana_logicalTRUE) yield significant but negative parameter estimates.

[1]The repository can by cloned to a local computer by entering in following command in a terminal (with Git installed):

```
git clone https://github.com/bquast/Making-Next-Billion-Demand-Access.git
```

In addition to this, the covariates included in the estimation are also highly significant. The highest education level of the individual (best_edu) and the household income (hhincome) are both positive and significant. The parameter estimate of a_womanTRUE here is negative but not at all significant, this is unsurprising as we use Internet expenditure at a household level. Most women live in a household which includes men and vise versa, suggesting that this effect cannot be isolated in this estimation. We further investigate this issue in a separate estimation discussed below. The variables describing linguistic skills in reading and writing in both English and the native language do yield many significant results, though lower levels of English writing skill seems to be correlated with a lower propensity to use the Internet (a_edlitwrten for levels 2 and 3, but not the very lowest: 4).

In an alternative formulation, we include the native language variable as a categorical variable (a_lng), interacted with the interface_intro dummy. In this estimation we only find significantly positive results for Setswana and Venda (as small language from the region bordering Zimbabwe), and a significantly negative effect for the language Afrikaans.

Table 1: Living in Household that Spent on Internet (last 30 days)

```
lm(h_nfnet ~  interface_intro*setswana_logical +
              factor(a_edlitrden)                    +
              factor(a_edlitwrten)                   +
              factor(a_edlitrdhm)                    +
              factor(a_edlitwrthm)                   +
              a_woman                                +
              hhincome                               +
              best_edu)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0008565	0.0018351	-0.4667340	0.6406924
interface_introTRUE	-0.0117376	0.0012170	-9.6449716	0.0000000
setswana_logicalTRUE	-0.0135286	0.0024255	-5.5777454	0.0000000
factor(a_edlitrden)2	0.0015879	0.0040272	0.3942904	0.6933684
factor(a_edlitrden)3	0.0002543	0.0052755	0.0481963	0.9615600
factor(a_edlitrden)4	-0.0036574	0.0068160	-0.5365861	0.5915561
factor(a_edlitwrten)2	-0.0101933	0.0040178	-2.5369998	0.0111839
factor(a_edlitwrten)3	-0.0107676	0.0052214	-2.0621916	0.0391951
factor(a_edlitwrten)4	-0.0071685	0.0066937	-1.0709421	0.2842010
factor(a_edlitrdhm)2	-0.0025418	0.0036848	-0.6898084	0.4903181
factor(a_edlitrdhm)3	-0.0028899	0.0051291	-0.5634293	0.5731453
factor(a_edlitrdhm)4	-0.0079926	0.0069070	-1.1571655	0.2472107
factor(a_edlitwrthm)2	0.0008117	0.0037229	0.2180373	0.8274010
factor(a_edlitwrthm)3	0.0013993	0.0051372	0.2723915	0.7853222
factor(a_edlitwrthm)4	-0.0069174	0.0069137	-1.0005354	0.3170567
a_womanTRUE	-0.0013881	0.0011452	-1.2121278	0.2254696
hhincome	0.0000027	0.0000001	46.2987087	0.0000000
best_edu	0.0013582	0.0001164	11.6710813	0.0000000
interface_introTRUE:setswana_logicalTRUE	0.0119717	0.0038666	3.0961443	0.0019617

Furthermore, we also use a variant of the base model, in which the propensity of adults (`a_owncom`) to own a computer is used as an explanandum. This is of particular relevance, as the explanandum here (`a_owncom`) differs from the base model's explanandum in two ways. Firstly, it does not include expenditure on Internet in ways such as Internet cafes, but focusses on actual ownership, signaling a more long-term investment and interest. Secondly, the `h_nfnet` variable is at a household level, whereas the `a_owncom` variable is at the level of an individual adult. The results from this estimation are included in Table 2. This form of the estimation yields similar results to those estimated in the base model. Firstly we find that the variable of interest, the interaction term between the event and the language dummy (`interface_introTRUE:setswana_logicalTRUE`) is positive and highly significant, with a p-value smaller than `0.001`. The individual dummy variables (`interface_introTRUE` and `setswana_logicalTRUE`) again are significant and negative with the former's p-value smaller than `0.01` and the latter's smaller than `0.001`. In terms of the linguistic skill, we find that the lower levels of English reading as well as English writing are correlated with lower propensities of computer ownership. Similar to Internet expenditure model, household income (`hhincome`) and highest level of education (`best_edu`) are both positive and highly significant (p-value: `~0`). However, unlike in the household Internet expenditure model, the sex of the individual here is highly significant, specifically, parameter estimate of a_womanTRUE is negative and highly significant (p-value: `~0`). As mentioned above, this variable is difficult to interpret when using a household-level variable as an explanandum, however, here, the computer ownership variable is at an individual level, which makes the coefficient more interpretable.

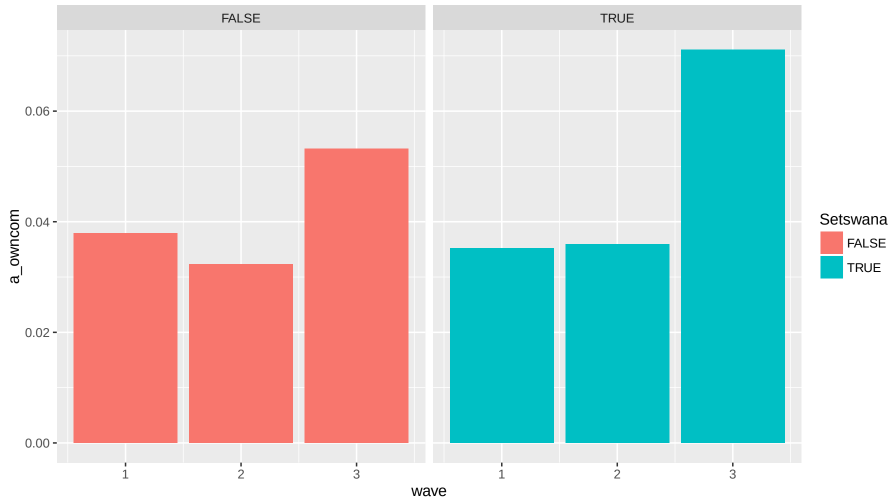
Table 2: Computer Ownership

```
lm(a_owncom ~ interface_intro*setswana_logical +
              factor(a_edlitrden)                    +
              factor(a_edlitwrten)                   +
              factor(a_edlitrdhm)                    +
              factor(a_edlitwrthm)                   +
              a_woman                                +
              hhincome                               +
              best_edu)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0075393	0.0031352	2.4047103	0.0161891
interface_introTRUE	-0.0053820	0.0020917	-2.5729824	0.0100856
setswana_logicalTRUE	-0.0147502	0.0041653	-3.5411916	0.0003987
factor(a_edlitrden)2	-0.0306749	0.0069077	-4.4406931	0.0000090
factor(a_edlitrden)3	-0.0310090	0.0090330	-3.4328617	0.0005978
factor(a_edlitrden)4	-0.0389174	0.0116674	-3.3355679	0.0008519
factor(a_edlitwrten)2	-0.0173238	0.0068925	-2.5134122	0.0119602
factor(a_edlitwrten)3	-0.0210640	0.0089384	-2.3565895	0.0184476
factor(a_edlitwrten)4	-0.0187291	0.0114540	-1.6351572	0.1020227
factor(a_edlitrdhm)2	-0.0017925	0.0063219	-0.2835349	0.7767681
factor(a_edlitrdhm)3	-0.0041655	0.0088001	-0.4733526	0.6359638
factor(a_edlitrdhm)4	-0.0267964	0.0118797	-2.2556443	0.0240974
factor(a_edlitwrthm)2	0.0010896	0.0063817	0.1707311	0.8644360
factor(a_edlitwrthm)3	-0.0020020	0.0088176	-0.2270503	0.8203856
factor(a_edlitwrthm)4	-0.0357886	0.0118921	-3.0094517	0.0026186
a_womanTRUE	-0.0230335	0.0019598	-11.7530143	0.0000000
hhincome	0.0000058	0.0000001	56.8794124	0.0000000
best_edu	0.0058419	0.0002002	29.1761771	0.0000000
interface_introTRUE:setswana_logicalTRUE	0.0238052	0.0066799	3.5637169	0.0003660

Figure 4: Computer Ownership Setswana

```
NIDS %>%
        group_by(Setswana = a_lng == 'Setswana', wave)   %>%
        filter(a_lng != 'Afrikaans')                      %>%
        summarise(a_owncom = mean(a_owncom, na.rm=TRUE)) %>%
        ggplot(aes(x=wave, y=a_owncom, fill=Setswana))   %+%
        geom_bar(stat='identity')                         %+%
        facet_grid(~Setswana)
```



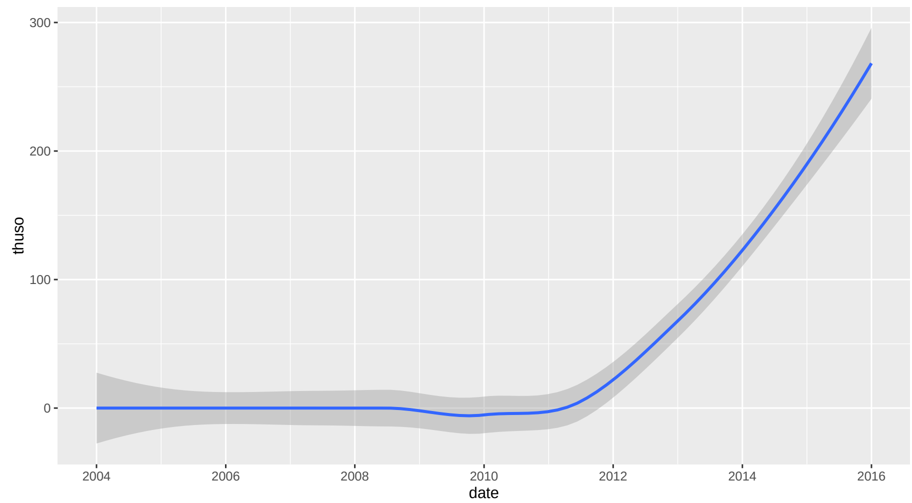
Asides from the effect on the proportion of households with Internet expenditure, and the proportion of adults who own a computer, we also estimate any possible effects on the probability of the household to spend on a cellphone, and on adults likelihood to own a cellphone.

We find no significant effects on the propensity of expenditure on cell phones or the propensity to own one. As discussed in the introduction, we suspect this to be a consequence of the fact that language switching on mobile cannot be automatic, since the Android operating system does not support the Setswana language, combined with the fact that the Setswana interface button is not visible directly on the `google.co.za` homepage, but rather in a dropdown menu (Figure 10).

This last effect is also visible in Figure 5, which illustrates how usage of the Setswana word ‘thuso’, meaning ‘help’ in search queries, was zero prior to the introduction of the Setswana search interface and became widespread thereafter.

Figure 5: Usage of Setswana Words on Google.co.za

```
ggplot(thuso) + geom_smooth(aes(x = date, y = thuso))
```



5 Conclusions and Limitations

In conclusion, despite recent advances in the reach, speed, and afordability of Internet connectivity in sub-Saharan Africa, actual connectivity has been stagnant.

This paper exploits the introduction of the Setswana-language interface on `google.co.za`, as a spillover of the development of this interface for `google.co.bw` and finds that it leads to a substantial increase in Internet usage and computer ownership among native Setswana speakers.

This increase of Internet usage among the Setswana speaking population as a result of the newly introduced interface language on `google.co.za`, suggests that there is a serious lack in the availability of local content in many African languages, which serves as an impediment to further Internet adoption here. Additionally, there was also an observably positive effect on computer ownership amongst adults. This suggests that the effect is unlikely to be ephemeral in nature, since computer ownership constitutes a more long-term investment in Internet access.

Finally, the fact that this effect is not observed in cellphone ownership and expenditure, which is possibly a consequence of the fact that the interface cannot automatically switch and that the link 'Setswana' is not directly visible from the landing page, suggests that it is important to make the link directly visible on the landing page in order to fully reach users.

References

Abadie, Alberto

2005 “Semiparametric difference-in-differences estimators”, *The Review of Economic Studies*, 72, 1, pp. 1-19.

Croissant, Yves

2013 “pglm: Panel Generalized Linear Model”, *R package version 0.1-2*, `http://CRAN.%20R-project.org/package=%20pglm`.

Croissant, Yves, Giovanni Millo, et al.

2008 “Panel data econometrics in R: The plm package”, *Journal of Statistical Software*, 27, 2, pp. 1-43.

Gandal, Neil

2006 “Native language and Internet usage”, *International journal of the sociology of language*, 2006, 182, pp. 25-40.

Git Team

2016 *Git: Software Code Manager*, 137 Montague ST STE 380, Brooklyn, NY 11201-3548, `http://www.git-scm.org/`.

Hoekwater, Taco, Hartmut Henkel, and Hans Hagen

2016 *LuaTeX*, `http://www.luatex.org/`.

Imbens, Guido W. and Jeffrey M. Wooldridge

2009 “Recent Developments in the Econometrics of Program Evaluation”, *Journal of Economic Literature*, 47, 1 (Mar. 2009), pp. 5-86, DOI: `10.1257/jel.47.1.5`, `http://www.aeaweb.org/articles/?doi=10.1257/jel.47.1.5`.

Jensen, Robert and Emily Oster

2009 “The power of TV: Cable television and women’s status in India”, *The Quarterly Journal of Economics*, pp. 1057-1094.

Knuth, Donald Ervin

1984 “Literate programming”, *The Computer Journal*, 27, 2, pp. 97-111.

Lamport, Leslie

1985 *I1 (\ LaTeX)—A Document*, pub-AW, vol. 410.

Leisch, Friedrich

2002 “Sweave: Dynamic generation of statistical reports using literate data analysis”, in *Compstat*, Springer, pp. 575-580.

LyX Team

2016 *LyX*, Free Software Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston, MA 02110-1301, USA, `http://www.lyx.org/`.

Otlogetswe, Thapelo J.

2010 “Setswana Google is here!”, *T.J. Otlogetswe Blog*, `http://otlogetswe.com/2010/08/13/setswana-google-here/`.

PeÃÂ±a-LÃÂ^3*pez, Ismael*

1999 "Challenges to the Network: Internet for Development".

R Core Team

2016 *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, `http://www.R-project.org/`.

Sanou, Brahima

2015 "The World in 2015: ICT facts and figures", *International Telecommunications Union*.

Sinai, Todd and Joel Waldfogel

2004 "Geography and the Internet: Is the Internet a Substitute or a Complement for Cities?", *Journal of Urban Economics*, 56, 1, pp. 1-24.

Southern Africa Labour and Development Research Unit

2008 *National Income Dynamics Study, Wave 1*, version 5.3, `http://www.nids.uct.ac.za/home/`.

2012 *National Income Dynamics Study, Wave 2*, version 2.3, `http://www.nids.uct.ac.za/home/`.

2013 *National Income Dynamics Study, Wave 3*, version 1.3, `http://www.nids.uct.ac.za/home/`.

Standage, Tom

2006 "Connecting the next billion", *The Economist-The World in 2006*, p. 117, `http://www.economist.com/node/5134746`.

Venables, William N and Brain D. Ripley

2013 *Modern applied statistics with S-PLUS*, Springer Science & Business Media.

Viard, V Brian and Nicholas Economides

2014 "The Effect of Content on Global Internet Adoption and the Global Ã¢ÂÂDigital DivideÃ¢ÂÂ", *Management Science*, 61, 3, pp. 665-687.

Xie, Yihui

2015 *Dynamic Documents with R and knitr*, Chapman and Hall/CRC, vol. 29, ISBN: 978-1498716963, `http://yihui.name/knitr/`.

Zeileis, Achim

2004 "Econometric Computing with HC and HAC Covariance Matrix Estimators", *Journal of Statistical Software*, 11, 10, pp. 1-17, `http://www.jstatsoft.org/v11/i10/`.

2006 "Object-Oriented Computation of Sandwich Estimators", *Journal of Statistical Software*, 16, 9, pp. 1-16, `http://www.jstatsoft.org/v16/i09/`.

Zeileis, Achim and Torsten Hothorn

2002 "Diagnostic Checking in Regression Relationships", *R News*, 2, 3, pp. 7-10, `http://CRAN.R-project.org/doc/Rnews/`.

6 Clustering

Table 3: Clustering

```
library(plm)     # panel linear model estimation
library(pglm)    # panel generalized linear model estimation
library(lmtest)  # Standard Error corrections
library(broom)   # output formatting using tidy()

# specify panel model
plm4_3 <- formula(as.numeric(h_nfnet)  ~ post_event*setswana  +
                                         factor(a_edlitrden)  +
                                         factor(a_edlitwrten) +
                                         factor(a_edlitrdhm)  +
                                         factor(a_edlitwrthm) +
                                         a_woman              +
                                         hhincome             +
                                         best_edu              )

# estimate
plm4_3e <- plm(plm4_3, data=pNIDS, model='within')

# correct errors
tidy( coeftest(plm4_3e, vcov=vcovHC(plm4_3e,
                                    type="HC0",
                                    cluster="group")) )
```

term	estimate	std.error	statistic	p.value
post_eventTRUE	-0.0019402	0.0012392	-1.5657830	0.1174144
setswanaTRUE	0.0027958	0.0137038	0.2040204	0.8383396
factor(a_edlitrden)2	0.0061294	0.0039180	1.5643993	0.1177388
factor(a_edlitrden)3	0.0028083	0.0042618	0.6589578	0.5099301
factor(a_edlitrden)4	0.0001637	0.0044802	0.0365363	0.9708551
factor(a_edlitwrten)2	-0.0088583	0.0039162	-2.2619681	0.0237095
factor(a_edlitwrten)3	-0.0080226	0.0042172	-1.9023800	0.0571351
factor(a_edlitwrten)4	-0.0060115	0.0046454	-1.2940715	0.1956549
factor(a_edlitrdhm)2	-0.0029617	0.0043786	-0.6764141	0.4987852
factor(a_edlitrdhm)3	-0.0035411	0.0061579	-0.5750552	0.5652601
factor(a_edlitrdhm)4	-0.0056185	0.0068075	-0.8253295	0.4091939
factor(a_edlitwrthm)2	0.0020764	0.0043717	0.4749531	0.6348253
factor(a_edlitwrthm)3	0.0045419	0.0063446	0.7158756	0.4740761
factor(a_edlitwrthm)4	0.0077089	0.0074611	1.0332094	0.3015178
a_womanTRUE	-0.0012994	0.0012422	-1.0461001	0.2955268
hhincome	0.0000007	0.0000003	2.4693445	0.0135439
best_edu	0.0001411	0.0003131	0.4507013	0.6522095
post_eventTRUE:setswanaTRUE	0.0070425	0.0034066	2.0673056	0.0387175

7 Dependent Variables Descriptive Statistics

The below table breaks down computer and cellphone ownership as well as Internet and cellphone expenditure by linguistic group.

Table 4: Descriptive statistics on Ownership and Expenditure

```
adulthh %>%
  group_by(a_lng, wave) %>%
  summarise(a_owncel    = mean(a_owncel,    na.rm = TRUE),
            a_owncom    = mean(a_owncom,    na.rm = TRUE),
            h_nfcel     = mean(h_nfcel,     na.rm = TRUE),
            h_nfnet     = mean(h_nfnet,     na.rm = TRUE))
```

a_lng	wave	a_owncel	a_owncom	h_nfcel	h_nfnet
IsiNdebele	1	0.6026490	0.0331126	0.6533333	0.0000000
IsiNdebele	2	0.6864865	0.0270270	0.6800000	0.0284091
IsiNdebele	3	0.7611111	0.0333333	0.8722222	0.0000000
IsiXhosa	1	0.4631115	0.0112269	0.4352518	0.0012043
IsiXhosa	2	0.6065066	0.0186335	0.5996664	0.0054517
IsiXhosa	3	0.7564988	0.0345508	0.7041694	0.0019756
IsiZulu	1	0.5610984	0.0132693	0.5756053	0.0013730
IsiZulu	2	0.5169004	0.0127744	0.4721318	0.0086366
IsiZulu	3	0.7662405	0.0252420	0.7247881	0.0044928
Sepedi	1	0.6127214	0.0265273	0.5772947	0.0048309
Sepedi	2	0.7029372	0.0226818	0.5933485	0.0021994
Sepedi	3	0.7936063	0.0718697	0.7863152	0.0050477
Sesotho	1	0.6562986	0.0366044	0.5507812	0.0062598
Sesotho	2	0.6973684	0.0457010	0.5411671	0.0213640
Sesotho	3	0.8114210	0.0949535	0.8147901	0.0113032
Setswana	1	0.5796003	0.0351724	0.5281593	0.0068681
Setswana	2	0.6004872	0.0359537	0.6494778	0.0037783
Setswana	3	0.7728036	0.0711086	0.7684564	0.0106264
SiSwati	1	0.6593060	0.0441640	0.7823344	0.0000000
SiSwati	2	0.7598870	0.0612813	0.6693227	0.0091185
SiSwati	3	0.8247978	0.0458221	0.7816712	0.0134771
Tshivenda	1	0.5980861	0.0334928	0.5645933	0.0000000
Tshivenda	2	0.7804878	0.0000000	0.9621212	0.5441176
Tshivenda	3	0.8617363	0.0225080	0.8456592	0.0000000
IsiTsonga	1	0.6411765	0.0235294	0.4408284	0.0000000
IsiTsonga	2	0.7375887	0.0118203	0.7163636	0.0929368
IsiTsonga	3	0.8621495	0.0397196	0.8691589	0.0023529
Afrikaans	1	0.5392884	0.1345441	0.6227876	0.0465116
Afrikaans	2	0.5422477	0.0904605	0.6258591	0.0424710
Afrikaans	3	0.6686971	0.1106225	0.7645862	0.0399458
English	1	0.7266667	0.2969374	0.7449933	0.1016043
English	2	0.7976190	0.3234127	0.8728814	0.1070707
English	3	0.8608059	0.3156934	0.8811700	0.1023766

8 General Descriptive Statistics

Figure 6: Household Income

```
ggplot(adulthh, aes(x=hhincome, fill=a_lng )) +
    stat_bin(bins=50)
```

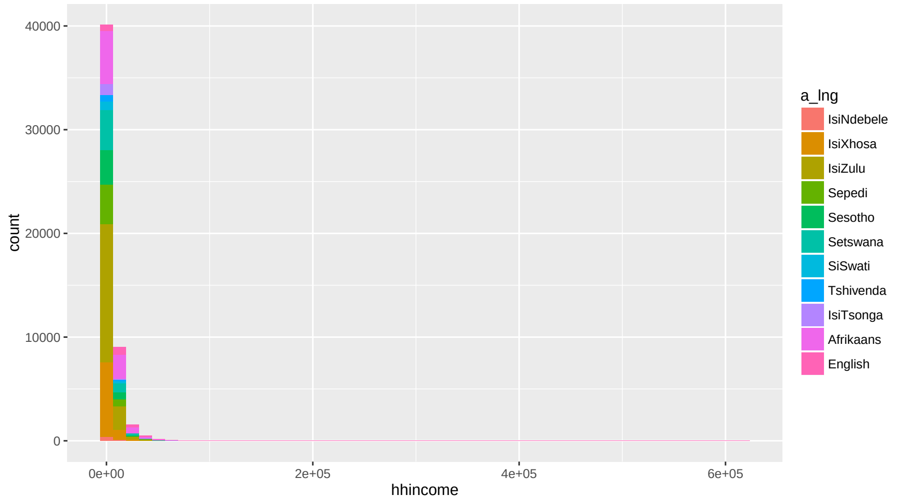
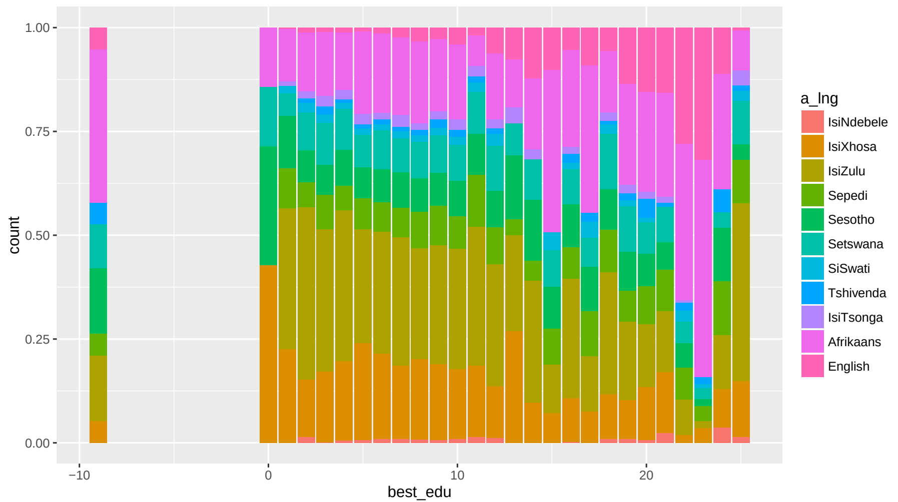


Figure 7: Years of Education

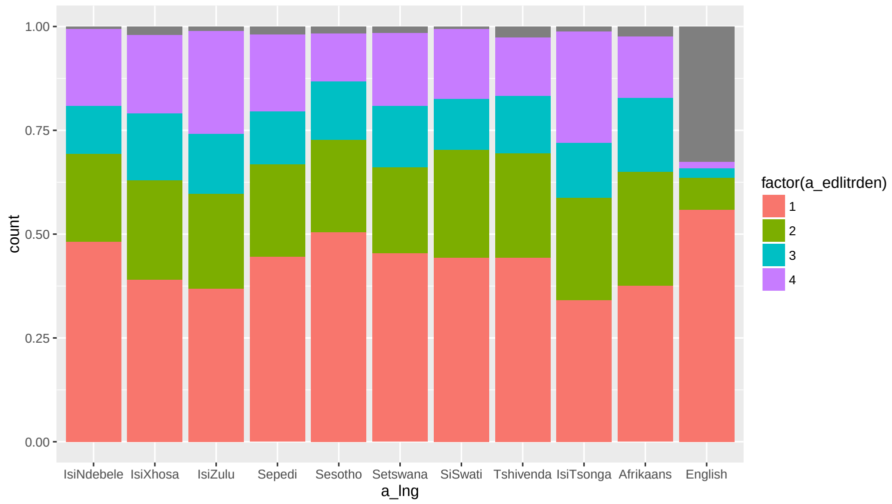
```
ggplot(adulthh, aes(x=best_edu, fill=a_lng )) +
    geom_bar(position = 'fill')
```



The Figure 8 describes the skill of individuals in reading the English language, where 1 the best and 4 is the worst, grey values are NA.

Figure 8: English Language Reading Skills

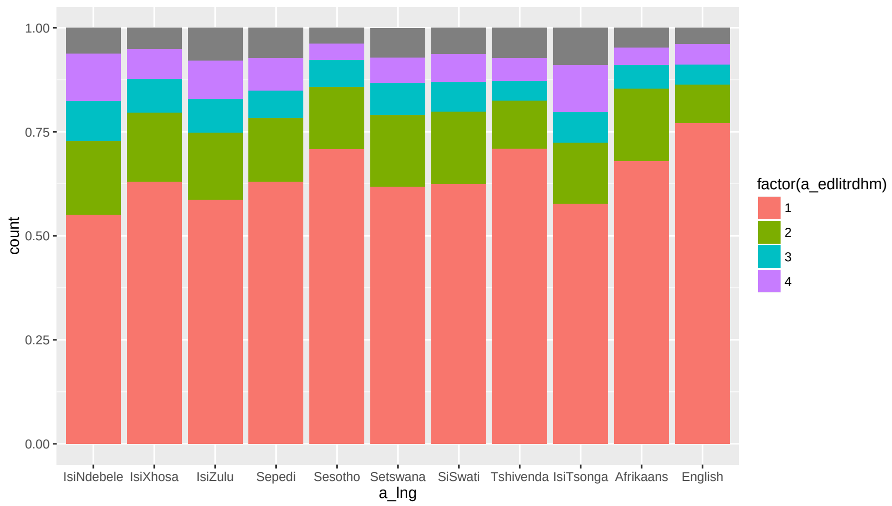
```
ggplot(adulthh, aes(x = a_lng, fill = factor(a_edlitrden)) ) +
        geom_bar(position = 'fill')
```



The Figure 9 does the same but with regards to the native language.

Figure 9: Native Language Reading Skills

```
ggplot(adulthh, aes(x = a_lng, fill = factor(a_edlitrdhm)) ) +
        geom_bar(position = 'fill')
```



9 Language Switching on Mobile

The graphics in Figure 10 illustrate the process of switching the mobile `google.co.za` interface language to Setswana, from the default English. The mobile interface language can automatically be changed based on the system language of the operation system (in most cases Android), however, as Setswana and many African languages are not available as system languages in Android, the website interface on `google.co.za` will default to English.

Figure 10: Changing Interface Language on Mobile