Making the 'Next Billion' Demand Access

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

Bastiaan Quast

Abstract

This paper shows that an exogenous increase in accessability of local language content leads to a very large increase in demand for internet connectivity among native speakers, even as demand as a whole is falling. Internet connectivity provides enormous improvements in quality of life as well as opportunities for the newly connected. Attempts to connect the 'next billion' in Africa have not met expectations, even in places where infrastructure has come online and prices have gone down. The introduction of the Setswana (Tswana) language in the South-African Google Search website was a spillover effect of this translation work being done for the Botswanan Google Search website. This exogenous event created a large increase in the number of internet-connected native speakers, as well as usage of the Setswana language online. Suggesting that connecting the fourth billion will require a greater focus on demand by mean of local content.

1 Introduction

Local content is a vital means to connecting new internet users.

Since the term 'Connecting the Next Billion' was introduced in The Economists 2006 'End of Year Report' (Standage, 2006), close to 2 billion people have been estimated to have been connected to the internet, up from the just over one billion people back then (Sanou, 2015). Yet despite increased range and improved affordability, many key growth markets such as sub-Saharan Africa are showing stagnation in the growth of internet connections.

This paper shows that exogenous increase in accessability of local content gave rise to a vast increase in the number of internet users 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 directly across the border in South Africa, where it is also one of the official state languages. This led to the introduction of the Setswana language on the South African Google website (`google.co.za`) as spillover of the translation work for Google's Botswanan website. This exogenous led to a vast increase in the number of native Setswana speakers reporting to have spent some amount of money in the past 30 days on internet access.

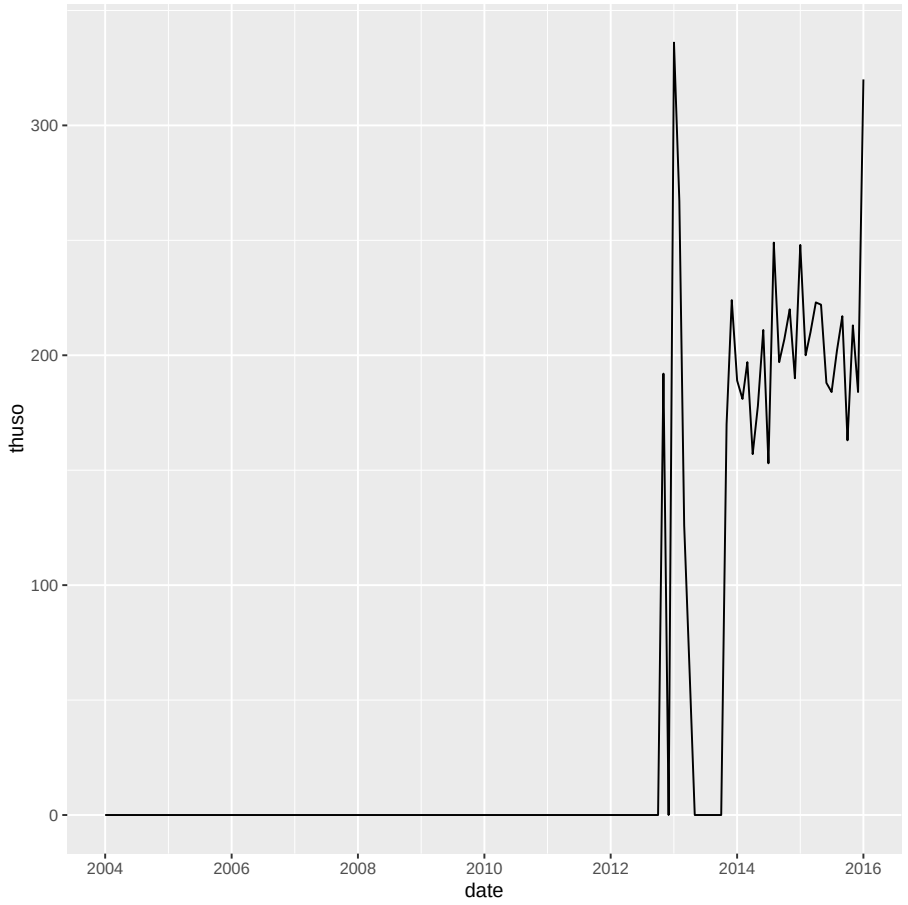
It is not required to use a certain interface language in order to search for content in this language. However, the search page is in many instances the first website viewed by users and the asides from being able to understand the interface, having the interface be in a certain language also encourages usage of this language, which in turn reveales more content in this language.

In short we can identify several major channels which promote further engagement, which together constitute the theory of change.

1. Being able to read and understand the words.
2. Encouragement from the familiarity with the language on what is often the first website visited.
3. Increased likelihood to using local language (see Figure 1) and thus finding more content in the native language.

Figure 1: Usage of Setswana Words on Google.co.za

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



The vast majority of internet access is developing countries is through handheld devices such as smartphones. However, due to the limited 'real estate' on a mobile website, the link to changing the interface language is replace with a dropdown menu that reveales the additional language options. Generally, the website will default the the operating system (Android / iOS) language, however, since many local African languages are not available as a system language, this is not possible there. The fact that the introduction seems to benefit deskop usage but not mobile usage is further substantiated by our results that isolate an increase in computer ownership, but no increase in cell phone ownership.

In addition to the increase in the number of individuals spending on an internet connection, we also find a positive effect on the number of individuals

living in households with a computer.

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 futher discussed in section 2. AFter whichsection 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-Difference estimator in subsection 3.2.

2 Data

- descriptive stats:
  - number of adults
  - Setswana speakers
  - people using internet / cell phone etc.
  - male / female
  - linguistic skills
  - income distribution

South Africa's National Income Dynamics Survey collect data on a representative set of around 10,000 households over time. The first survey took place in 2008, the second one in late 2010 and early 2011, and third one took place in 2012 (Southern Africa Labour and Development Research Unit, 2008, 2012, 2013).

3 Methods

This section beings 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 allow us the rule out any possible endogeity issues that might

otherwise arise in context 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 Estimation

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-Difference estimator using a native-Setswana speaker dummy variable, interacted with a event dummy variable on the introduction of the Setswana interface on `google.co.za`.

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 once 'base' level, which is where all language dummies are `FALSE` (`0`) and the level (native language) thus has to be the nth one (here `IsiNdebele`).

The dependent variable here is also a dummy variable, which would normally allow for the usage of a estimator such as logit. However, since we are employing the Difference-in-Difference methodology ....

3.3 Software

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 in

the document is performed using the knitr implemenation (Xie, 2015) of the sweave framework (Leisch, 2002).

All changes are logged using the version control system Git (Git Team, 2016).

4 Results

In the base model, we use an interaction of the `post_event` dummy and `setswana` 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).

In an alternative formulation, we include the native language variable as a categorical variable, interacted with the `post_event` dummy. In this estimation we only find significantly positive results for `setswana` and `venda` (as small language from the north-eastern region), and a significantly negative effect for the language `afrikaans`.

Futhermore, we also use the base model with the propensity of adults (`a_owncom`) to own a computer as an explanandum, whereby we find similar results. This is of significance, because the explanandum is different from the base model's in two ways. Firstly, it does not include expenditure on internet in ways such as internet cafe. Secondly, the `h_nfnet` variable is at a household level, whereas the `a_owncom` variable is at an individual adult's level.

We find no significant effect on the 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 2).

```
lm(h_nfnet ~  post_event*setswana  +
              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)	-0.0007802	0.0018382	-0.4244139	0.6712660
post_eventTRUE	-0.0118598	0.0012192	-9.7275870	0.0000000
setswanaTRUE	-0.0136327	0.0024304	-5.6091357	0.0000000
factor(a_edlitrden)2	0.0015400	0.0040296	0.3821762	0.7023324
factor(a_edlitrden)3	0.0002169	0.0052826	0.0410653	0.9672440
factor(a_edlitrden)4	-0.0037000	0.0068226	-0.5423219	0.5875994
factor(a_edlitwrten)2	-0.0102695	0.0040201	-2.5545159	0.0106367
factor(a_edlitwrten)3	-0.0108074	0.0052286	-2.0669776	0.0387418
factor(a_edlitwrten)4	-0.0071782	0.0067009	-1.0712332	0.2840702
factor(a_edlitrdhm)2	-0.0026154	0.0036909	-0.7086027	0.4785746
factor(a_edlitrdhm)3	-0.0029346	0.0051387	-0.5710741	0.5679522
factor(a_edlitrdhm)4	-0.0080255	0.0069159	-1.1604527	0.2458705
factor(a_edlitwrthm)2	0.0008491	0.0037294	0.2276776	0.8198979
factor(a_edlitwrthm)3	0.0013947	0.0051466	0.2709891	0.7864006
factor(a_edlitwrthm)4	-0.0069746	0.0069204	-1.0078396	0.3135367
a_womanTRUE	-0.0014132	0.0011472	-1.2318972	0.2179937
hhincome	0.0000028	0.0000001	46.4115514	0.0000000
best_edu	0.0013633	0.0001165	11.6980667	0.0000000
post_eventTRUE:setswanaTRUE	0.0120675	0.0038748	3.1143779	0.0018445

Table 1: Computer in Household

```
lm(a_owncom ~ post_event*setswana  +
              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)	0.0076236	0.0031357	2.4312519	0.0150504
post_eventTRUE	-0.0054318	0.0020923	-2.5960394	0.0094334
setswanaTRUE	-0.0147962	0.0041674	-3.5504730	0.0003849
factor(a_edlitrden)2	-0.0306508	0.0069013	-4.4412904	0.0000090
factor(a_edlitrden)3	-0.0309578	0.0090312	-3.4278696	0.0006089
factor(a_edlitrden)4	-0.0389988	0.0116608	-3.3444228	0.0008252
factor(a_edlitwrten)2	-0.0174611	0.0068860	-2.5357494	0.0112239
factor(a_edlitwrten)3	-0.0210480	0.0089368	-2.3552002	0.0185168
factor(a_edlitwrten)4	-0.0187070	0.0114489	-1.6339589	0.1022741
factor(a_edlitrdhm)2	-0.0018694	0.0063225	-0.2956791	0.7674765
factor(a_edlitrdhm)3	-0.0042389	0.0088029	-0.4815352	0.6301384
factor(a_edlitrdhm)4	-0.0267295	0.0118766	-2.2506064	0.0244150
factor(a_edlitwrthm)2	0.0012267	0.0063829	0.1921868	0.8475967
factor(a_edlitwrthm)3	-0.0019980	0.0088200	-0.2265280	0.8207918
factor(a_edlitwrthm)4	-0.0357502	0.0118852	-3.0079574	0.0026315
a_womanTRUE	-0.0229898	0.0019601	-11.7288993	0.0000000
hhincome	0.0000058	0.0000001	56.9305683	0.0000000
best_edu	0.0058348	0.0002002	29.1431892	0.0000000
post_eventTRUE:setswanaTRUE	0.0238541	0.0066835	3.5690970	0.0003586

5 Conclusions and Limitations

The vast increase of internet usage among the Setswana speaking population as a result of the newly introduced interface language on `goog.co.za`, suggest 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.

References

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/`.

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/`.

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*.

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`.

Xie, Yihui

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

A Ownership and Expenditure by Native Language

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

Table 2: Descriptive statistics on Ownership and Expenditure

```
adulthh %>%
  group_by(a_lng, wave) %>%
  summarise(owncel = mean(a_owncel,   na.rm = TRUE),
                          owncom = mean(a_owncom,   na.rm = TRUE),
                          cel    = mean(h_nfcel,    na.rm = TRUE),
                          net    = mean(h_nfnet,    na.rm = TRUE),
                          celspn = mean(h_nfcelspn, na.rm = TRUE),
                          netspn = mean(h_nfnetspn, na.rm = TRUE))
```

a_lng	wave	owncel	owncom	cel	net	celspn	netspn
1	1	0.6026490	0.0331126	0.6533333	0.0000000	59.78519	0.0000000
1	2	0.6864865	0.0270270	0.6800000	0.0284091	84.12571	1.4204545
1	3	0.7611111	0.0333333	0.8722222	0.0000000	105.46067	0.0000000
2	1	0.4631115	0.0112269	0.4352518	0.0012043	39.76659	0.0574297
2	2	0.6065066	0.0186335	0.5996664	0.0054517	54.47248	0.2978972
2	3	0.7564988	0.0345508	0.7041694	0.0019756	86.33322	0.0869279
3	1	0.5610984	0.0132693	0.5756053	0.0013730	48.21557	0.0363844
3	2	0.5169004	0.0127744	0.4721318	0.0086366	49.00402	1.3053671
3	3	0.7662405	0.0252420	0.7247881	0.0044928	96.59072	0.6889696
4	1	0.6127214	0.0265273	0.5772947	0.0048309	40.47351	0.3462158
4	2	0.7029372	0.0226818	0.5933485	0.0021994	51.69161	0.1583578
4	3	0.7936063	0.0718697	0.7863152	0.0050477	110.64450	1.1385306
5	1	0.6562986	0.0366044	0.5507812	0.0062598	49.45647	0.4772727
5	2	0.6973684	0.0457010	0.5411671	0.0213640	60.29976	3.0657354
5	3	0.8114210	0.0949535	0.8147901	0.0113032	115.52993	1.8284574
6	1	0.5796003	0.0351724	0.5281593	0.0068681	59.33404	0.2886598
6	2	0.6004872	0.0359537	0.6494778	0.0037783	86.17885	1.0214106
6	3	0.7728036	0.0711086	0.7684564	0.0106264	112.96165	0.8509804
7	1	0.6593060	0.0441640	0.7823344	0.0000000	59.17647	0.0000000
7	2	0.7598870	0.0612813	0.6693227	0.0091185	63.28685	0.4589666
7	3	0.8247978	0.0458221	0.7816712	0.0134771	120.19944	1.4555256
8	1	0.5980861	0.0334928	0.5645933	0.0000000	108.45631	0.0000000
8	2	0.7804878	0.0000000	0.9621212	0.5441176	50.85606	1.2058824
8	3	0.8617363	0.0225080	0.8456592	0.0000000	142.95035	0.0000000
9	1	0.6411765	0.0235294	0.4408284	0.0000000	27.80896	0.0000000
9	2	0.7375887	0.0118203	0.7163636	0.0929368	55.88364	0.9368030
9	3	0.8621495	0.0397196	0.8691589	0.0023529	103.93128	0.3529412
10	1	0.5392884	0.1345441	0.6227876	0.0465116	133.43521	9.4163569
10	2	0.5422477	0.0904605	0.6258591	0.0424710	103.81572	10.9746890
10	3	0.6686971	0.1106225	0.7645862	0.0399458	146.19560	10.5008501
11	1	0.7266667	0.2969374	0.7449933	0.1016043	371.01291	31.0356653
11	2	0.7976190	0.3234127	0.8728814	0.1070707	375.17797	35.3555556
11	3	0.8608059	0.3156934	0.8811700	0.1023766	377.87127	23.3816514
12	1	0.8214286	0.1428571	0.7407407	0.1481481	241.74074	23.4444444
12	2	0.7000000	0.1111111	0.9000000	0.0000000	263.00000	0.0000000
NA	1	0.6434783	0.0695652	0.5546392	0.0301783	115.03832	4.2275242
NA	2	0.6029412	0.0147059	0.5893720	0.0429936	113.54589	7.5764331
NA	3	0.9000000	0.6000000	0.7609756	0.0337349	170.08854	4.0556901

B Factor vs. Dummy

In addition to estimating our model using a dummy variable for native Setswana speakers, we also estimate the model using a factor variable of the categorical variable describing language. In a linear model this is employed as a dummy variable for each level except for the base level. The results of this estimation are similar to the base model, suggesting that the results are robust to specification idiosyncrasies.

Table 3: Factor of Language

```
lm(h_nfnet ~  post_event*factor(a_lng))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0153374	0.0069104	2.2194845	0.0264583
post_eventTRUE	-0.0153374	0.0116070	-1.3213959	0.1863755
factor(a_lng)2	-0.0119771	0.0071295	-1.6799254	0.0929782
factor(a_lng)3	-0.0101387	0.0070313	-1.4419371	0.1493265
factor(a_lng)4	-0.0118839	0.0073298	-1.6212958	0.1049606
factor(a_lng)5	-0.0017102	0.0073480	-0.2327407	0.8159637
factor(a_lng)6	-0.0100812	0.0072710	-1.3864957	0.1656019
factor(a_lng)7	-0.0106935	0.0084765	-1.2615403	0.2071202
factor(a_lng)8	0.1182366	0.0101957	11.5966572	0.0000000
factor(a_lng)9	0.0258487	0.0085674	3.0171200	0.0025532
factor(a_lng)10	0.0293054	0.0071303	4.1099620	0.0000396
factor(a_lng)11	0.0884438	0.0077638	11.3917636	0.0000000
factor(a_lng)12	0.0927707	0.0216448	4.2860579	0.0000182
post_eventTRUE:factor(a_lng)2	0.0139527	0.0119551	1.1670893	0.2431800
post_eventTRUE:factor(a_lng)3	0.0146315	0.0117940	1.2405899	0.2147632
post_eventTRUE:factor(a_lng)4	0.0169315	0.0122240	1.3851020	0.1660275
post_eventTRUE:factor(a_lng)5	0.0130134	0.0123009	1.0579188	0.2900977
post_eventTRUE:factor(a_lng)6	0.0207076	0.0121878	1.6990391	0.0893181
post_eventTRUE:factor(a_lng)7	0.0241705	0.0141697	1.7057877	0.0880539
post_eventTRUE:factor(a_lng)8	-0.1182366	0.0155235	-7.6166264	0.0000000
post_eventTRUE:factor(a_lng)9	-0.0234958	0.0140356	-1.6740145	0.0941341
post_eventTRUE:factor(a_lng)10	0.0106404	0.0119616	0.8895436	0.3737153
post_eventTRUE:factor(a_lng)11	0.0139328	0.0132554	1.0511063	0.2932149

Lastly, we also estimate the factor model with the inclusion of the linguistic skill variables. These results are again similar to the ones from using simply a dummy variable for native Setswana speakers, suggesting robustness to specification idiosyncrasies.

Table 4: LM4_1: with read / write in eng / native and woman

```
lm(h_nfnet ~  post_event*factor(a_lng) +
              a_edlitrden             +
              a_edlitwrten            +
              a_edlitrdhm             +
              a_edlitwrthm            +
              a_woman)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0318595	0.0071928	4.4293548	0.0000095
post_eventTRUE	-0.0160615	0.0115788	-1.3871439	0.1654045
factor(a_lng)2	-0.0099108	0.0073031	-1.3570604	0.1747686
factor(a_lng)3	-0.0074446	0.0072117	-1.0322986	0.3019376
factor(a_lng)4	-0.0117912	0.0075144	-1.5691408	0.1166220
factor(a_lng)5	-0.0024480	0.0075095	-0.3259828	0.7444388
factor(a_lng)6	-0.0090931	0.0074556	-1.2196300	0.2226114
factor(a_lng)7	-0.0109606	0.0086869	-1.2617360	0.2070501
factor(a_lng)8	0.1112536	0.0102699	10.8329397	0.0000000
factor(a_lng)9	0.0241967	0.0088206	2.7432030	0.0060866
factor(a_lng)10	0.0327904	0.0073039	4.4894292	0.0000072
factor(a_lng)11	0.0832838	0.0079063	10.5339061	0.0000000
factor(a_lng)12	0.0929623	0.0216230	4.2992237	0.0000172
a_edlitrden	-0.0033726	0.0022402	-1.5054902	0.1322049
a_edlitwrten	-0.0068192	0.0021992	-3.1007081	0.0019317
a_edlitrdhm	0.0005543	0.0021514	0.2576646	0.7966669
a_edlitwrthm	0.0021000	0.0021548	0.9745694	0.3297790
a_womanTRUE	-0.0012292	0.0011597	-1.0599272	0.2891832
post_eventTRUE:factor(a_lng)2	0.0132434	0.0119238	1.1106677	0.2667171
post_eventTRUE:factor(a_lng)3	0.0145833	0.0117679	1.2392479	0.2152600
post_eventTRUE:factor(a_lng)4	0.0177728	0.0121993	1.4568644	0.1451606
post_eventTRUE:factor(a_lng)5	0.0129743	0.0122610	1.0581779	0.2899798
post_eventTRUE:factor(a_lng)6	0.0200726	0.0121631	1.6502852	0.0988914
post_eventTRUE:factor(a_lng)7	0.0253515	0.0141395	1.7929538	0.0729868
post_eventTRUE:factor(a_lng)8	-0.1117654	0.0154157	-7.2500872	0.0000000
post_eventTRUE:factor(a_lng)9	-0.0181736	0.0140391	-1.2944988	0.1954996
post_eventTRUE:factor(a_lng)10	0.0095912	0.0119282	0.8040750	0.4213578

C Language Switching on Mobile

Figure 2: Changing Interface Language on Mobile