Examining Web User Flows and Behaviours in CLARIN Ecosystem

Go Sugimoto (Austrian Centre for Digital Humanities)
CLARIN Annual Conference 2017, Budapest, Hungary 2017-09-20
Agenda

1. Motivation
2. Method
3. Statistics & Interpretation
4. Conclusion
1. Number Game 2.0.x

We now have some ideas about the user traffic. It is a good step forward for the evaluation of CLARIN. Don’t let it die there. Move on!
Previous studies

- Eckart et al. (2015)
- Sugimoto (2016)

Analysis only for
- a single website
- a technical website (=VLO)
Web Analytics overview
2. Method - Measuring CLARIN ecosystem

Point of Entry (PoE)

User

CLARIN main

CLARIN applications

CLARIN partners
Comparative study (Traffic in and out of):

- Clarin.eu (main website)
- Virtual Language Observatory (VLO)
- Weblicht
- Content Search Aggregation
- Identity Provider
- Discovery Service

1 February 2016 - 31 January 2017
Where you were BEFORE

Standpoint

Where you were AFTER
ICAME corpora in CLARIN

ICAME corpora have found a new home in the CLARIN project through the Norwegian part, CLARINO.

The corpora are available for academic use.

The corpora are available in the Corpuscle program (login from the top line).

Users from many universities can log in with their ordinary user id through the eduGAIN or CLARIN Service Provider Federation (SPF). Go to the Corpuscle home page and choose eduGAIN or CLARIN SPF from the top login line and search for your university.

Norwegian users can use Feide login through eduGAIN (if you don’t find your institution on the login page contact Knut.Hofland@uni.no).

If you are not able to use eduGAIN or CLARIN SPF register for an ClarinIdP account and be manually approved.

OpenIdP is another choice, but this may be terminated on short notice.

At the moment the following corpora are available for searching (most also for downloading through the "Overview" option in the menu to the left for each corpus):

- The Brown family (Brown, LOB, FLOB, Frown, BLOB and BE06) prepared at Lancaster (BLOB and BE06 not for downloading)
- FLOB and Frown with original POS tagging
- ACE (Australian Corpus of English)
- COLT (Corpus of London Teenage Language)
- Helsinki Corpus of English Texts
- Helsinki Corpus of Older Scots
- Helsinki CEECS (Corpus of Early English Correspondence Sampler)
- London-Lund Corpus

5.3% of referrers

Outgoing Traffic

51% Internal pages
12% Services
11% Event
8.8% Consortia
5.5% Nutshell
5% Users (probably log-in)

2.8% Outlinks
30% VLO
3.8% CLARIN-D

Some noises

40% Exit

Depends on what is put (or promoted) on the home page which changes frequently
Do they know CLARIN already?

100% - (Reload and Internal pages) = 14817 (external access)
Search engine with CLARIN + Direct Entry = 8028 (existing users?)

= 54.2% minimum + other channels
<table>
<thead>
<tr>
<th>URL</th>
<th>Unique Clicks</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>vlo.clarin.eu</td>
<td>998</td>
<td>11.2%</td>
</tr>
<tr>
<td>hdl.handle.net</td>
<td>615</td>
<td>6.9%</td>
</tr>
<tr>
<td>centres.clarin.eu</td>
<td>439</td>
<td>4.9%</td>
</tr>
<tr>
<td>catalog.clarin.eu</td>
<td>404</td>
<td>4.5%</td>
</tr>
<tr>
<td><a href="http://www.clarin.eu">www.clarin.eu</a></td>
<td>386</td>
<td>4.3%</td>
</tr>
<tr>
<td>infra.clarin.eu</td>
<td>255</td>
<td>2.9%</td>
</tr>
<tr>
<td>lindat.mff.cuni.cz</td>
<td>227</td>
<td>2.6%</td>
</tr>
<tr>
<td><a href="http://www.clarin-d.de">www.clarin-d.de</a></td>
<td>215</td>
<td>2.4%</td>
</tr>
<tr>
<td>docs.google.com</td>
<td>181</td>
<td>2.0%</td>
</tr>
<tr>
<td>weblicht.sfs.uni-tuebingen.de</td>
<td>118</td>
<td>1.3%</td>
</tr>
</tbody>
</table>

Top 10 outlinks from the CLARIN main website
User Flow Plug-in (€79/year) is more powerful than Transition Reports. It can check the position a page was viewed, how often and how users navigated through your entire website. Multiple-user paths can be tracked.
At VLO home page

Incoming traffic
- 394 from internal pages
- 637 from internal searches
- 243 from search engines
- 1,102 from websites
- 1,477 direct entries

Outgoing traffic
- 1,589 to internal pages
- 1,218 internal searches
- 23 outlinks
- 1,074 exits

5,056 pageviews

From Websites
- clarin.eu/content/virtual-language-observatory 54%
- linguistics.stackexchange.com/questions/14982/looking-for-korean-text-corpus 7.4%
- clarin.eu/content/virtual-language-observatory-vlo 5.8%
- moodle.univie.ac.at/course/view.php?id=48590&section=5 3.2%
- moodle.univie.ac.at/course/view.php?id=53375&section=5 1.8%
- Others 27%

Direct Entries 29%

From Search Engines
- 4.8%

From Internal Search
- 22%

Internal Searches 24%
<table>
<thead>
<tr>
<th>Ranking</th>
<th>URL</th>
<th>Unique Clicks</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>hdl.handle.net</td>
<td>297</td>
<td>41.4%</td>
</tr>
<tr>
<td>2</td>
<td>infra.clarin.eu</td>
<td>118</td>
<td>16.4%</td>
</tr>
<tr>
<td>3</td>
<td><a href="http://www.clarin.eu">www.clarin.eu</a></td>
<td>96</td>
<td>13.4%</td>
</tr>
<tr>
<td>4</td>
<td>clarinportal.informatik.uni-leipzig.de</td>
<td>20</td>
<td>2.8%</td>
</tr>
<tr>
<td>5</td>
<td><a href="http://www.sil.org">www.sil.org</a></td>
<td>15</td>
<td>2.1%</td>
</tr>
<tr>
<td>6</td>
<td>clarin-pl.eu</td>
<td>12</td>
<td>1.7%</td>
</tr>
<tr>
<td>7</td>
<td>urn.fi</td>
<td>11</td>
<td>1.5%</td>
</tr>
<tr>
<td>8</td>
<td>diglib.hab.de</td>
<td>10</td>
<td>1.4%</td>
</tr>
<tr>
<td>9</td>
<td>clarinws.informatik.uni-leipzig.de:8080</td>
<td>8</td>
<td>1.1%</td>
</tr>
<tr>
<td>10</td>
<td>metashare.ut.ee</td>
<td>8</td>
<td>1.1%</td>
</tr>
</tbody>
</table>

Top 10 outlinks from the VLO
WEBLICHT - TYPES OF INCOMING TRAFFIC

- Direct Entry, 560, 29%
- Websites, 1330, 70%
- Search Engines, 17, 1%
WEBLICHT - WEB REFERERS

idp.clarin.eu 29%

idp.uni-tuebingen.de 9%

mylogin.ub.unifreiburg.de 7%

idp.uni-stuttgart.de 7%

idp.rrz.uni-koeln.de 4%

Others 44%
Limitations and Dirty Jobs for Analyses

- Duplicates, flattening, and development with CMS
  [https://www.clarin.eu/content/factsheet-clarin-plus](https://www.clarin.eu/content/factsheet-clarin-plus) (287 visits)
  [https://www.clarin.eu/node/4213](https://www.clarin.eu/node/4213) (235 visits)
- Handle PIDs have no clue for the content, while URLs include clues (domain names etc)
4. Conclusion

- Piwik (transition view) - simple yet powerful tool for user assessment
- CLARIN - complex web ecosystem with many PoE
- Good starting point for the better understanding of the users
- Tricky to analyse & interpret data. Don’t use default data blindly (as most people do...)
- Open Evaluation! **Agile** Evaluation!
Future work

• Combine with other surveys (questionnaire etc)
• Use/buy plug-ins for better analyses
• Web Analytics to synch with Outreach, User Involvement & Dev team
• Case study with the institutions in the consortium (1. Piwik data exchange to track the user flows at the partner site. 2. Distinguish the CLARIN users by IP address from non-CLARIN users etc)
• Social media analytics

Your ideas & collaboration welcome!
Köszönöm nagyon szépen!

Original title of this paper: “Donau in Blue” was abandoned with regret. The concept was:

- **User flow** = Danube river
- The Blue Danube (Strauss II) + Rhapsody in Blue (Gershwin)
- Vienna to Budapest

Go.Sugimoto@oeaw.ac.at
Extra information
1.4% of referrers. No link! (website changed? noise?)

http://www.soooperthemes.com/content/about -> http://clarin.eu
Searches related to clarin

diario clarin de hoy ultimas noticias
diario nacion
diario perfil
diario infobae
clarin empleos
diario cronica
clarin clasificados
pagina 12

clarin language lab software
clarin language definition
stanford digital language lab
language lab software for schools

digital language resource

● Search Engine Optimisation (SEO)!
● Existing users not new users? (likely)
● Are they CLARIN members? (likely)
VLO pageviews, outlinks, downloads

Pageviews
Downloads
Outlinks
More questions than answers?

Is CLARIN addressing
• the community of CLARIN, linguistics, (Digital) Humanities, or Cultural Heritage, etc?
• Europe or also outside Europe?
• Researchers, (higher) Education, Citizen (science) etc?

Did the users find and use information (in general, resource, tool)? Why do the users come and go?