The INFRASTRUKTUR scheme, The Research Council of Norway

User Survey Report

Project number 208375

Coordinator: Prof. Koenraad De Smedt, UiB
June 14, 2020

Background

A user survey of research infrastructures has been commissioned by the Research Council of Norway in its letter of March 23, 2020 (ref 20/35, Herman Farbrot).

The present report covers CLARINO (Common Language Resources and Technology Infrastructure Norway), supported by the RCN during 2012–2021 under contract 208375. The total budget of the infrastructure was 49.8 M NOK and the contribution from the RCN was 25 M NOK.

CLARINO is a national research infrastructure which forms part of the European CLARIN ERIC. The infrastructure has many international users, which have also been targeted in this user survey (see also Appendix). For this reason, the survey as well as this report are in English.

CLARINO is a distributed infrastructure with operative centres at The University of Bergen (UiB), The University of Oslo (UiO), The Arctic University of Norway (UiT) and the National Library of Norway (NB). These centers offer a range of data and services which are described elsewhere, most recently in the Fact Sheet which was also commissioned by the RCN, and which therefore will not be repeated here.

Among the resources and services to which CLARINO provides access are also those constructed in INESS (2010–2017, RCN grant 195323); these are also addressed in the present user survey.

This user survey covers the whole period during which the CLARINO and INESS infrastructures have been operating until the date of this report.

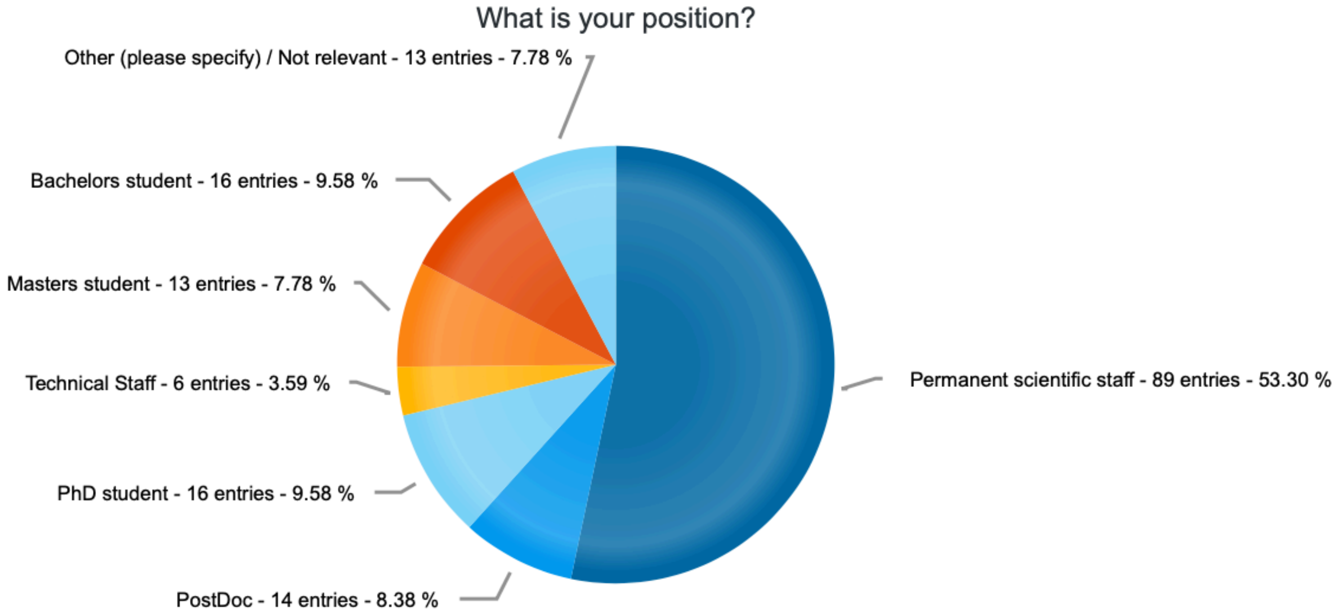
Target group and respondents

On May 27, 2020, the survey was sent to 3296 unique email addresses of identified users that were collected by the CLARINO centers. These are users who have logged in as required to access data with restricted licenses or to use personalized services. The actual number of CLARINO users is approximately twice as large when including users who have been counted but not identified because login is not necessary for accessing open data, in line with the OECD recommendations to make access unhindered and as easy as possible.

Consequently, the group addressed by this survey makes up about half of the estimated total number of users, and it might involve a slight bias towards those data and services which are not fully open. In the list of email addresses, about 1300 seem to be associated with a student domain.

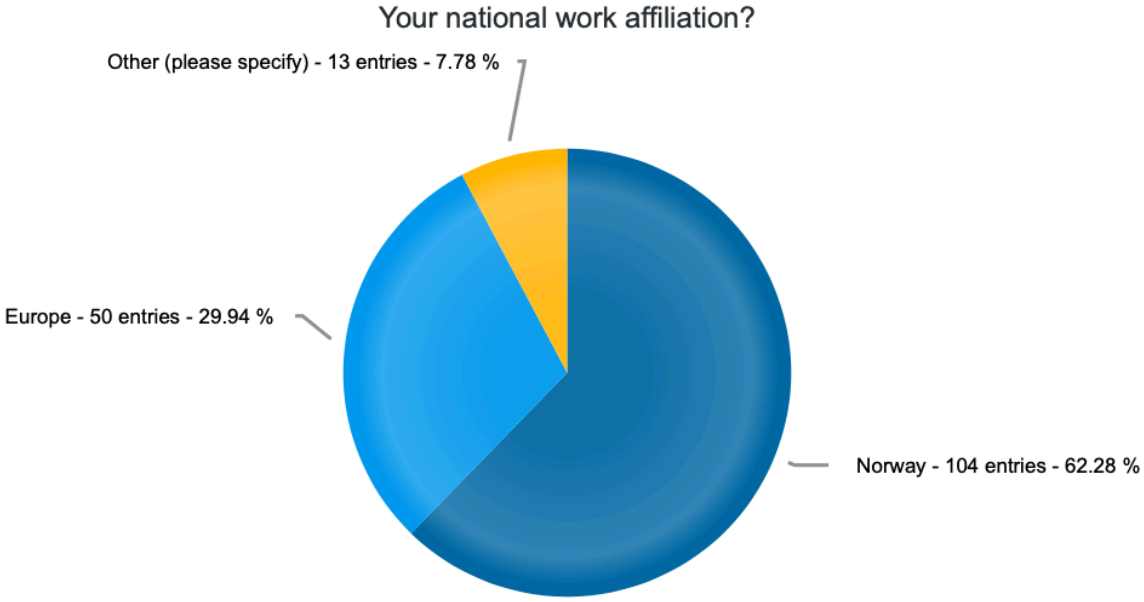
167 responses were received by June 14, 2020, when the survey was closed. Many email addresses failed, including all in the domains *stud.hioa.no* and *stud.hihm.no*. It is uncertain to which degree the received responses are representative of the whole population of CLARINO users. The respondents answered on behalf of themselves.

A breakdown of reported users’ positions is in the following chart.



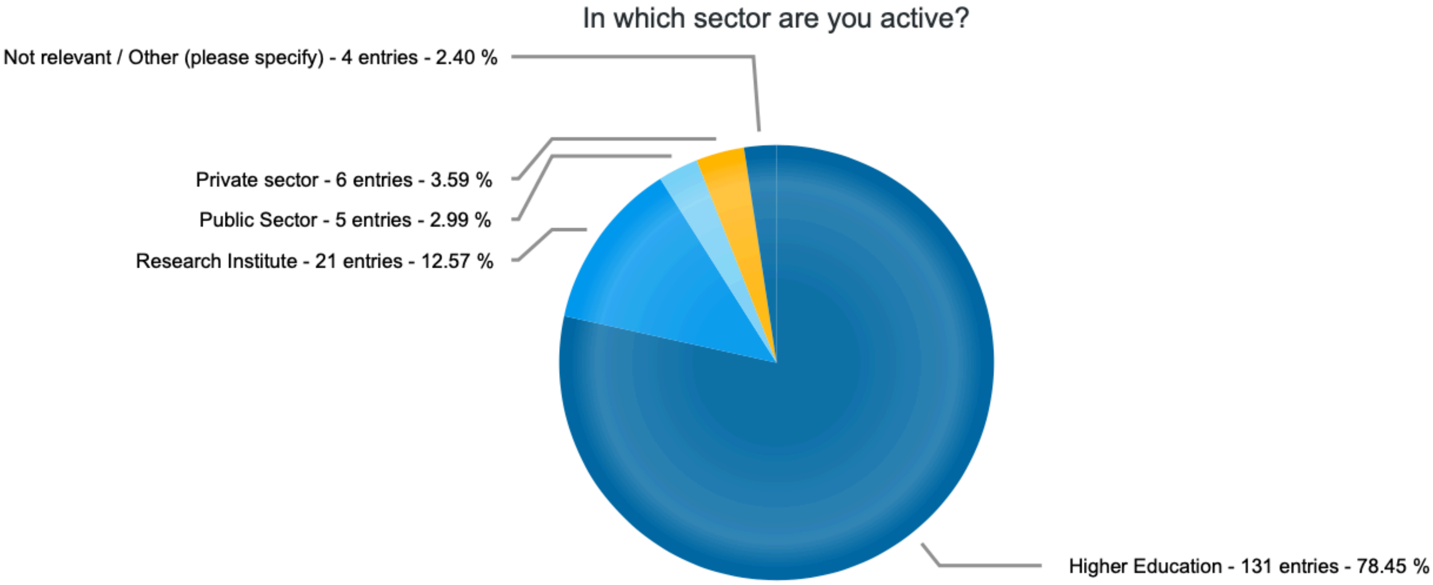
Permanent scientific staff is by far the largest group, covering more than half of the users. Students are about equally distributed over all levels (bachelors, masters, PhD) and together make up more than a quarter of the respondents. Other positions mentioned are civil servant, emeritus, lexicographer (2), librarian, non-permanent scientific staff, professor, professor emeritus, research and administrative positions, researcher (2), senior adviser, and temporary scientific staff.

A breakdown of national affiliation is in the following chart.



Users in Norway make up more than 62% but also the rest of Europe is a sizable part. Other answers are Australia, China, Faroe Islands, Japan, Russia, Switzerland (2) and USA (5).

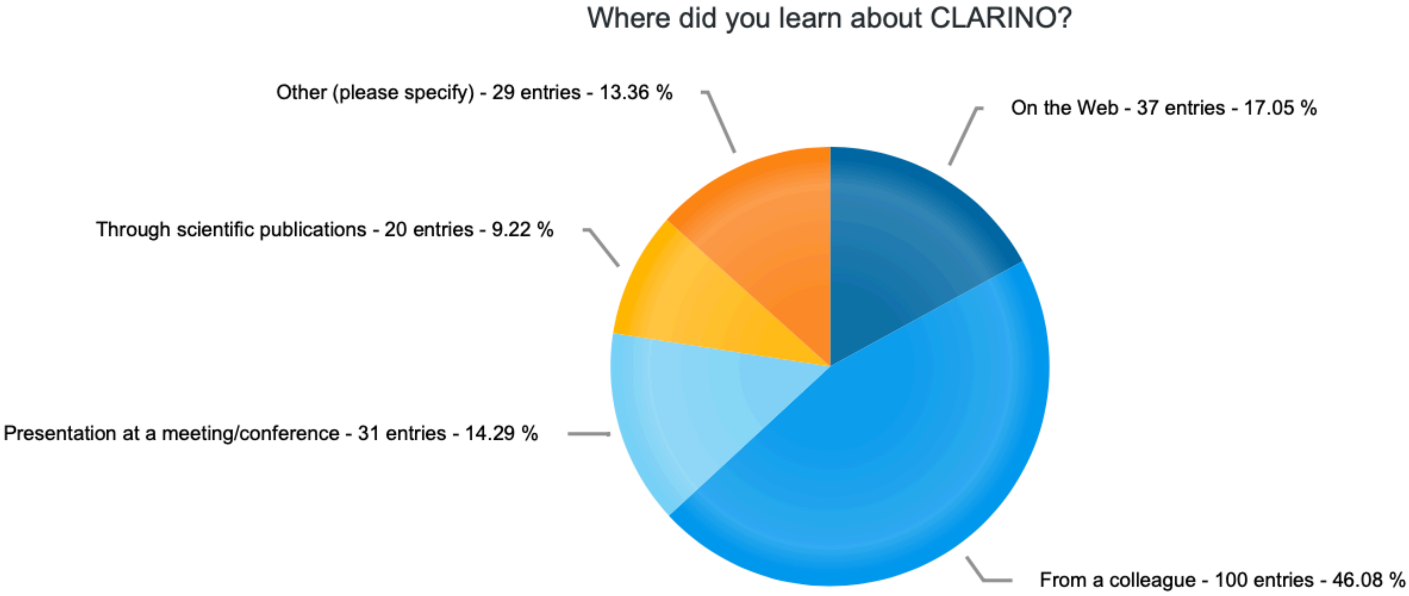
A breakdown according to sector is given in the following chart.



The higher education sector makes up more than three quarters of the users, followed by research institutes. The public and private sectors have few users. Other answers are unemployed, student and retired.

First acquaintance with the infrastructure

The following chart summarizes where users became aware of CLARINO.

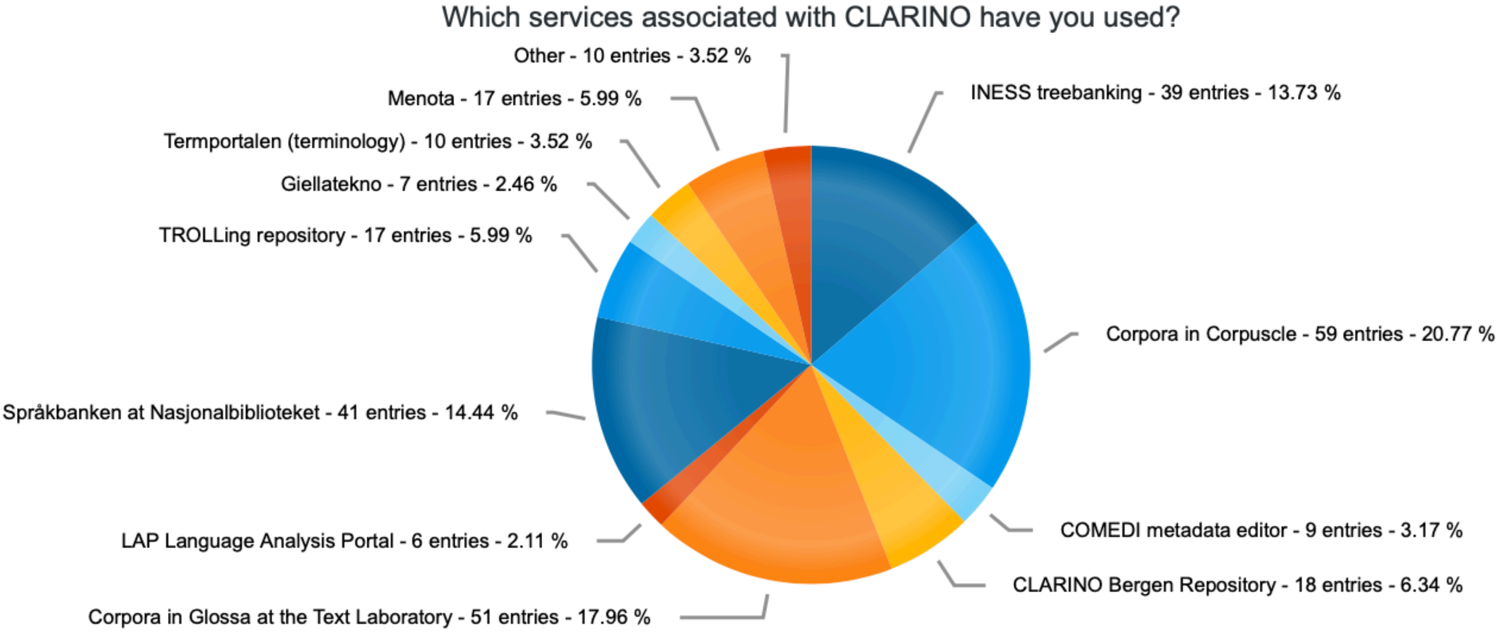


Communication with colleagues is the largest source of information, but conferences, publications and web-based presence are also important. Other answers are: At the University of Bergen Library, at university, CLARIN ERIC, co-creator, Cooperates with the Text Laboratory, Univ. of Oslo, From a lecturer, from my professor, Have not heard about it prior to survey email, I don’t know, I have never used it, I honestly have no idea of what it is, I work at a CLARINO-affiliated institution, I've not heard of it, In a class, In a university class, in class, In NLP class, Lecturer, My Bachelor's supervisor, other, past experience, Personal contact, contributed to various resources, previous work at linguistic project, Recommended by two of my professors, Through employment (both at CLARINO and related projects), Through previous employment, through university courses, Trial and error, work as a scientific researcher.

A few respondents do not seem to know CLARINO or claim to not have used it, even if their email address has been recorded by CLARINO centers. It is all the more strange because the instructions mention that the survey is addressed only at CLARINO users and “If you have never accessed any of the services mentioned in the survey, please disregard this request [to fill out the survey].”

Services used

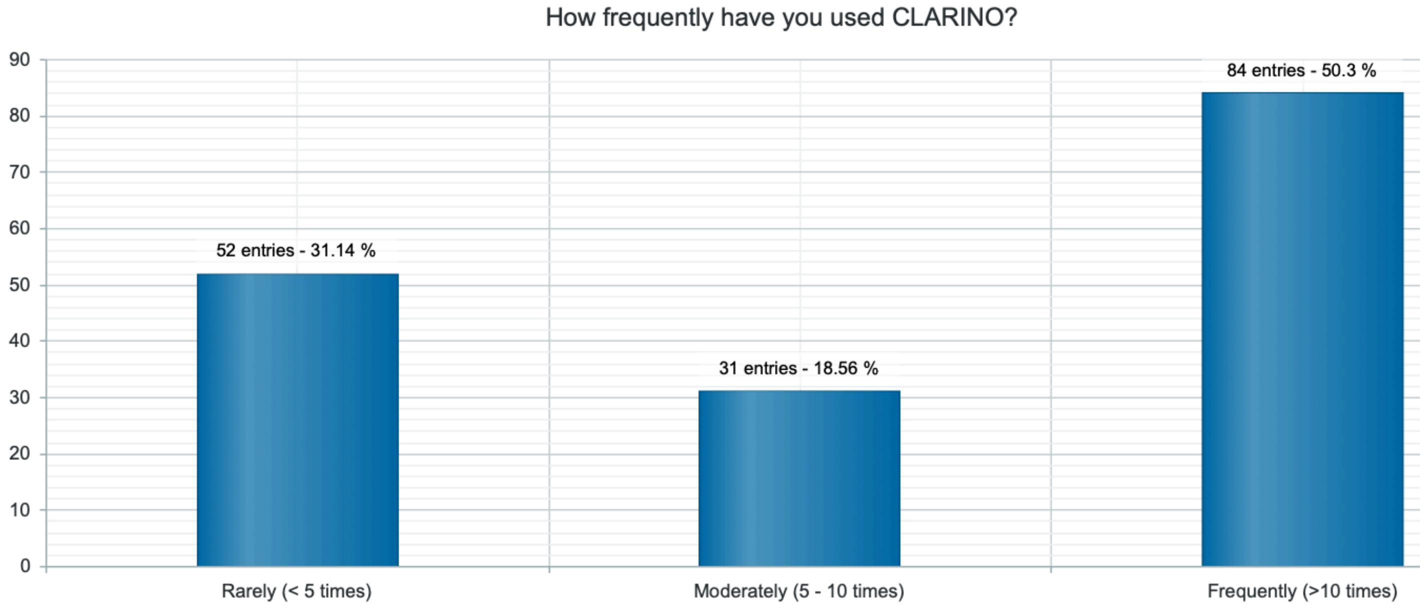
Since CLARINO is a distributed infrastructure with several centers, an important question is which services have been used. Each respondent can tick off several choices. The following figure provides a overview.



The service with the largest use is corpus access at Corpuscle (UiB), the corpora at the Text Laboratory (UiO), Språkbanken (NB) and INESS (UiB), but further named services together also make up a sizable proportion. Other answers are Aviskorpus, Glossa, I’ve porbably used som but don’t really know which, Menotec, None, Nordic Dialect Corpus, Talespråkskorpus at UiO, Talko, Uncertain, Wittgenstein Archives at the University of Bergen. A couple of these “other” answers are in fact subsumed by resources and services listed in the question.

Frequency of use

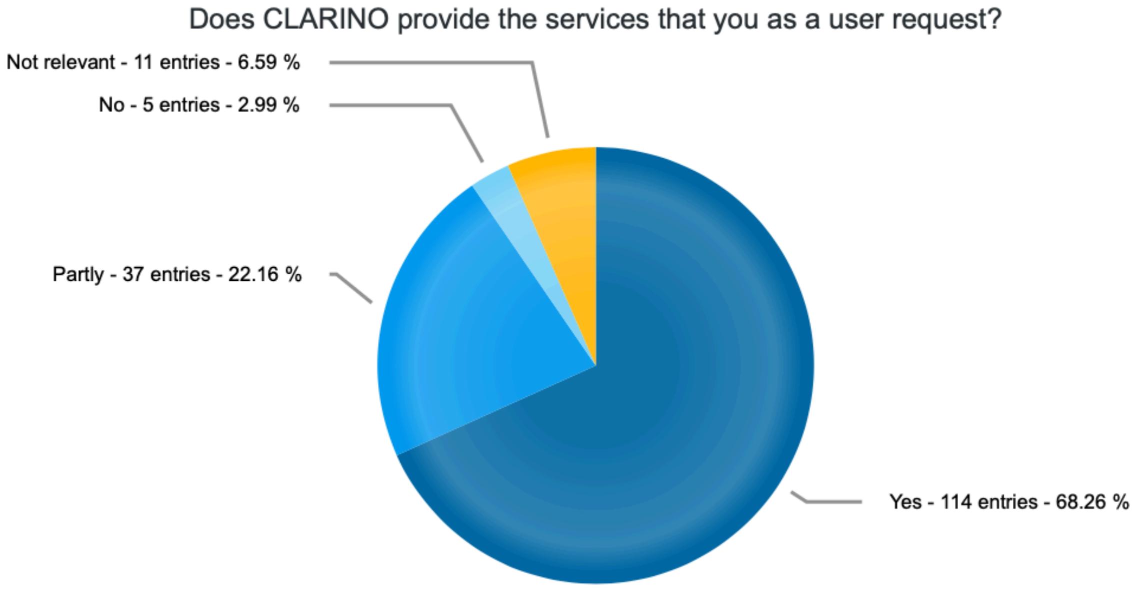
The following bar chart divides users according to the frequency of their use.



From these results it appears that CLARINO has a regular user base, but also caters to occasional users.

Satisfaction with types of services

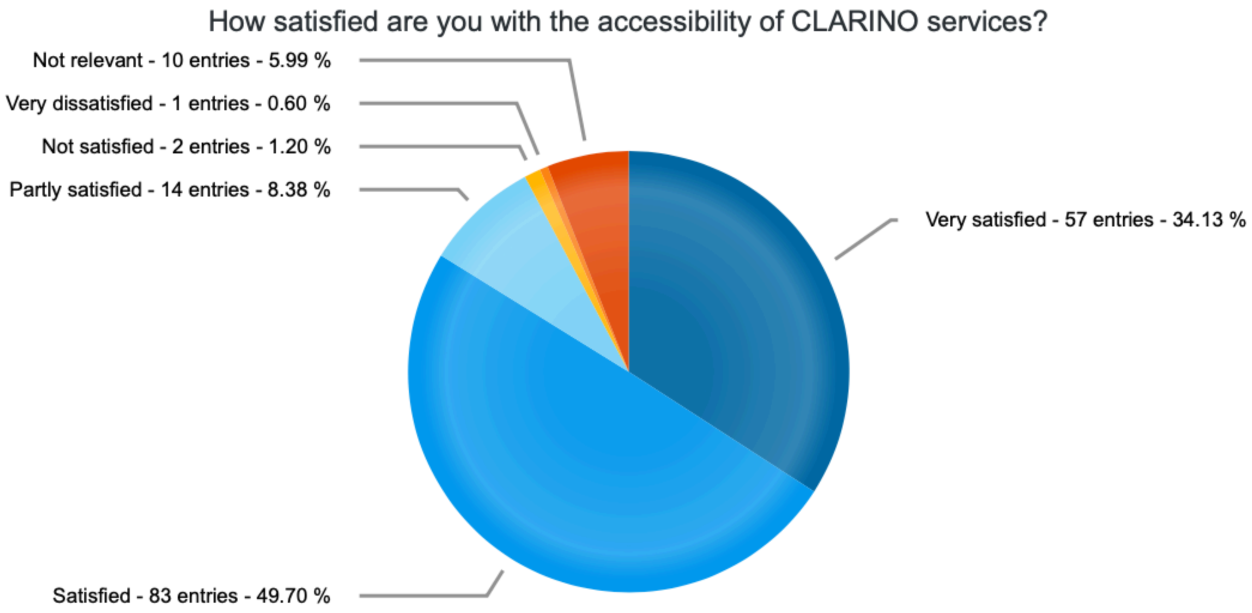
The following chart shows the degree of satisfaction with the provided services in relation to user's needs.



It may be concluded that users are largely satisfied with the services that are provided. It must be kept in mind that CLARINO does not itself construct new data, but aims at making existing digital language data more accessible for research and development. Even if CLARINO cannot cater to all possible user needs, it is remarkable that the answers are as positive as they are.

Satisfaction with accessibility of CLARINO services

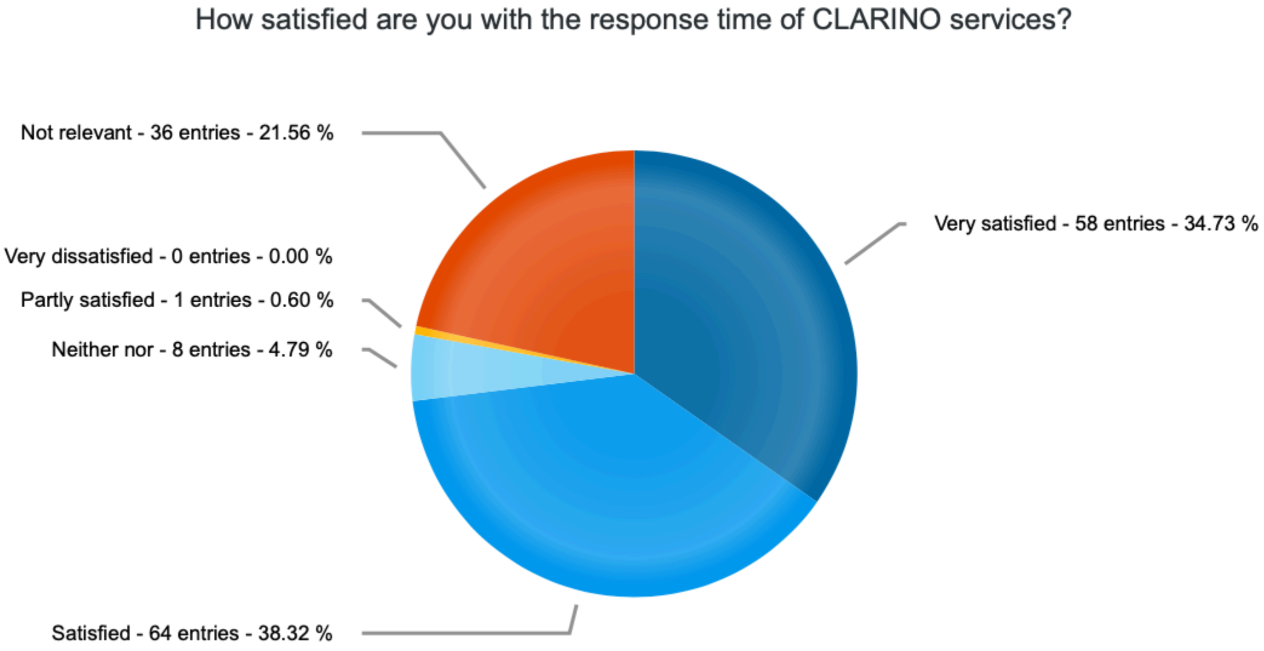
The following chart shows the degree of satisfaction with the accessibility of services.



It may be concluded that users are largely satisfied to very satisfied. All services are provided online.

Satisfaction with response time of CLARINO services

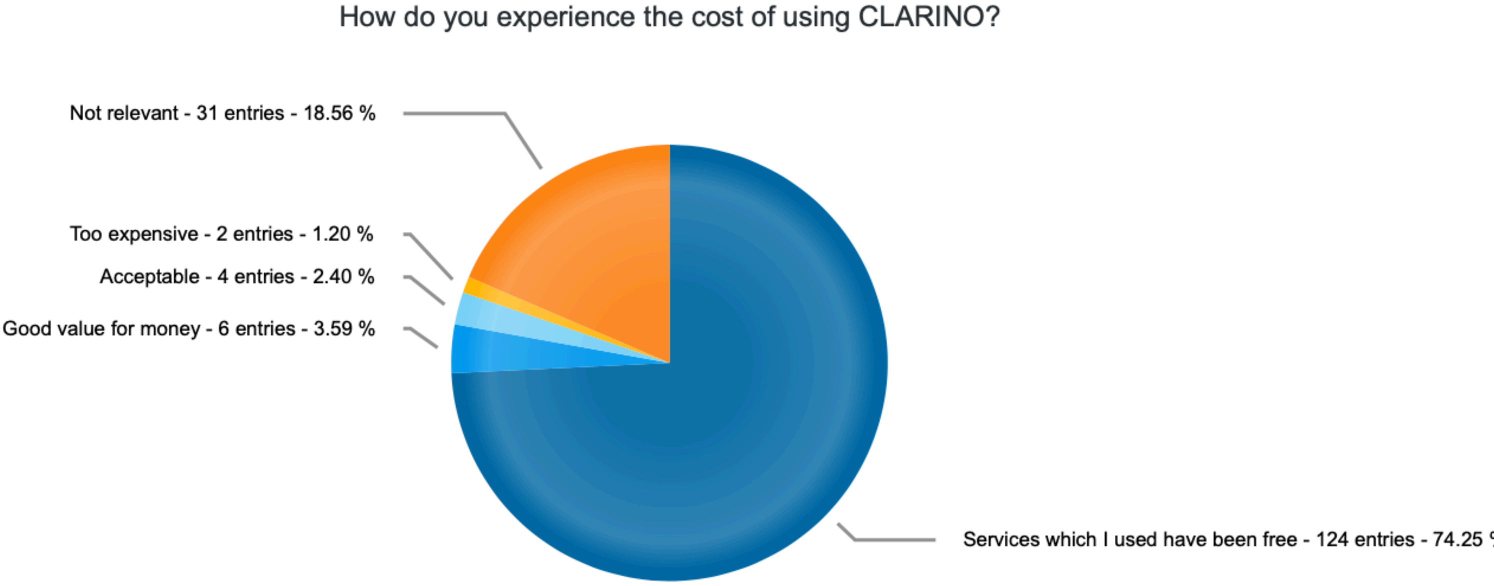
The following chart shows the degree of satisfaction with the accessibility of services.



It may be concluded that users are largely satisfied to very satisfied. It can be added that the uptime of CLARINO services is very high, in part thanks to investments in good computer systems.

Satisfaction with cost

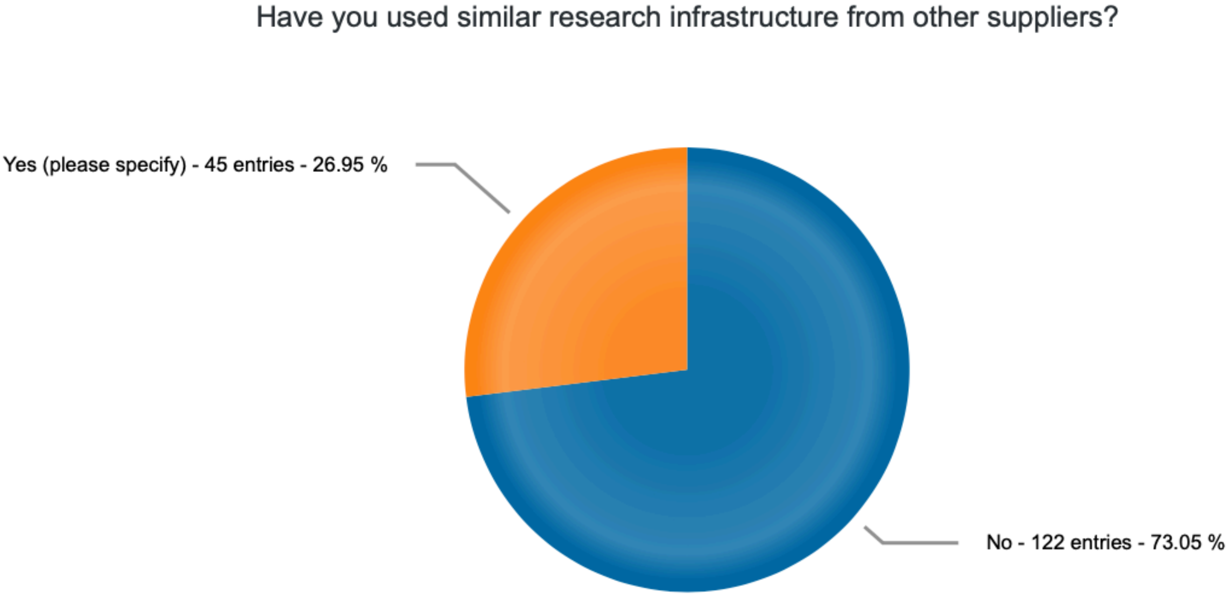
The following chart shows the degree of satisfaction with the cost of services.



Most CLARINO online services are free – in that case there is nothing to complain about. A few services that require manual assistance are priced according to the full cost (TDI) model, according to host institution policies, and this pricing can be regarded as not cheap. CLARINO must be careful not to price itself out of the market.

Use of other infrastructures

The following chart shows whether users have used other infrastructures services besides CLARINO.



Although almost three quarters of the replies are from users who do not report any user of other infrastructure services besides CLARINO, there are still 45 replies from users who do. Many of these respondents mention use of services by CLARIN ERIC nodes outside of Norway, in particular the Swedish and Finnish CLARIN nodes.

Suggestions for improvement of CLARINO

The following suggestions for improvement were given by respondents.

1. Å kunne få tilgang til NB sine ressurser
2. A regular (but short) meeting once a year/term in order to update and clarify status and progress.
3. All CLARINO resources should provide a clear user license which is clearly visible/accessible for users. Preferably international licenses, and with so few restrictions as possible. Preferably CC0 with guidance on how to cite. For instance, I cannot find any license information on the different webpages of Leksikografisk bokmålskorpus. So I don't really know what I'm allowed to do with material obtained from that corpus.
4. anglicisms in newspapercorpora, clearer demarcations
5. Automatic evaluation of requests for downloading corpora (rather than manual inspection)
6. I am missing a Norwegian social media chat corpus, but the other corpora at Corpuscel are great
7. I found INESS a bit difficult to search in, but with guidance from CLARINO working group members I was able to find exactly what I needed. Thank you!
8. I had hoped that the corpora in the Bergen collection would be made simply publicly available.
9. I have only used a few of the services and it has been a while since I last used them, so I do not have specific suggestions at this point.
10. I only heard about these services a few weeks ago through a university course on translation (French-Norwegian) at Østfold University College
11. I use the XLE web parser for English grammar course. I demonstrate syntactic analysis to my students. I wish there was a simpler version that can be used to demonstrate pedagogical grammar to students of grammar and linguistics at college level. This could have motivated students to learn grammar and understand it.
12. Improve quality of services.
13. Just keep up the good work
14. Many others should use Glossa as well.
15. More available help for analysing text material for corpora
16. More relevant materials
17. More training in use, more Norwegian corpus data (bokmål and nynorsk text), more terminological resources,
18. none
19. search acrosss several corpora
20. The Corpuscle code is relatively heavy, non-intuitive and complex, compared fex with glossa.
21. The search tool should be able to let you search for letters inside of words as well, not just complete words. I was looking for words that contained rn, but could not do a search on just those letters.
22. TROLLing is perfect.

CLARINO has the following comments on some of the above suggestions.

Ad 1. NB is independently working on this. The result may or may not become available through CLARINO. If CLARINO were to control, manage and provide access to every language resource in Norway, it would need more than a tenfold budget and considerable research policy pressure.

Ad 3. CLARINO aims at providing a license for all resources, but it takes time. Also, there seems to be a false assumption that CLARINO can decide on license types, whereas in fact this is controlled by the numerous different rightsholders.

Ad 4 and 21. There may be a false expectation that CLARINO can provide a direct answer to every research question. Instead, CLARINO provides access to resources but researchers must put in some effort as well.

Ad 5. Some licenses unfortunately require manual permission by resource owners based on research plans. There is a false expectation that this could be automated.

Ad 7, 11 and 20. Some search facilities and other tools are more complex than others because they offer a lot more options, as required by advanced users.

Ad 12 and 16. It is unclear what is requested. Users have too high expectations and do not realize that CLARINO does not commission the development of new language resources, but aims at improving access to existing digital resources.

Ad 19. This has been implemented in INESS and in Glossa (the latter through Federated Content Search), but it must be kept in mind that search across different resources can only be done if they are of the same data types.

Final remarks

From the survey responses it appears that CLARINO has satisfied to very satified users, although a few have unrealistic expectations from an infrastructure with low funding that provides most of its data for free. It also appears that some people do not realize what CLARINO is. A lesson may be that better information to the target user group is at least as important as maintaining and improving the infrastructure services.

Appendix

The targeted user group's top-level domains, giving an indication of the home locations and sectors of international identified users, are listed as follows, with their frequencies.

```
2276 no               3 tw
 159 fi               3 ee
 142 com              3 cl
 139 de               2 za
 122 se               2 lv
  64 nl               2 kr
  64 edu              2 hu
  33 lt               2 hk
  26 uk               2 fo
  24 dk               1 tr
  23 pl               1 sk
  23 fr               1 rs
  23 ch               1 org
  20 it               1 nz
  20 be               1 io
  18 bg               1 il
  15 cz               1 ge
  14 is               1 by
   9 ru               1 ax
   7 eu
   6 at
   5 jp
   5 es
   5 cn
   5 ca
   5 br
   5 au
   4 si
   4 gr
```