Daan Broeder and Jan Odijk

Sustainability and Genericity of CLARIN Services in the Netherlands

Abstract: Based on the ten years that have elapsed since the start of the CLARIN-NL project and its follow-up CLARIAH-NL, this chapter offers an analysis of the sustainability and genericity of services created in the context of CLARIN in the Netherlands. Our focus is on search applications, for which we make a proposal for coming to a more efficient and sustainable approach not only in the Netherlands but also CLARIN-wide. We also offer a number of general recommendations for improving sustainability of infrastructure services.

Keywords: sustainability of software services, genericity of services, specificity of services, research infrastructures, CLARIN, CLARIAH-NL

1 Introduction

In this chapter we analyse the sustainability and (lack of) genericity of services created in the context of CLARIN in the Netherlands. We interpret sustainability as the ability of (a set of) services to endure[1] over time. This goes beyond the sustainability of the service software and importantly also includes the aspects of being able to provide and manage cost-effective hosting and providing funds for the services' maintenance.

By service genericity we mean the aspect of a service being targeted at a broad number of tasks instead of focussing on one specific task only (specificity). Services created for (a limited number) of specific tasks are ideally maximally optimized for those tasks and adhere to the philosophy "do a few specific tasks

1 This is an extension of what is mentioned in Daniel S. Katz's blog on Software Sustainability https://danielskatzblog.wordpress.com/2016/09/13/defining-software-sustainability/.

Acknowledgements: We would like to thank colleagues who commented on parts of earlier versions of this chapter, in particular Katrien Depuydt, Jesse de Does, Jan Niestadt, and Vincent Vandeghinste (all from the Institute for the Dutch Language) as well as anonymous reviewers of an earlier version of this chapter.

Daan Broeder, CLARIN ERIC / Utrecht University, Utrecht, the Netherlands, e-mail: d.g.broeder@uu.nl
Jan Odijk, Utrecht University, Utrecht, the Netherlands, e-mail: j.odijk@uu.nl

https://doi.org/10.1515/9783110767377-006

very well". Although it is not impossible for generic tools to do many tasks very well, in practice this requires significant efforts and is expensive. Finding the optimal compromise between service genericity and specificity certainly is one important aspect of a service's sustainability. More than ten years have passed since the start of CLARIN-NL and the follow-up project CLARIAH-NL and we are now able to analyse and reflect on both issues, which are clearly interrelated. We will argue that a large number of search services developed in these projects are too specific and are better replaced by fewer but more generic search services in order to improve not only their sustainability but also the functionality they offer. All services mentioned offer a reference to extensive descriptions in the CLAPOP portal,[2] which also offers an overview of all NL CLARIN and CLARIAH[3] services via the CLAPOP search service.[4]

This chapter is structured as follows: First we present an overview on how the NL CLARIN infrastructure was populated with tools and services (Section 2). Subsequently, we present an overview of the different types of services thus obtained and an analysis of the different circumstances that determine their sustainability (Section 3). We then focus on the important sub-group of search applications, zooming in on the text search applications, for which we argue that their high specificity or lack of genericity leads to less sustainability and less functionality and propose an approach towards a more sustainable, more efficient way to manage the development and operation of the NL CLARIN search applications (Section 4). At the end of the chapter we conclude with a number of general observations and recommendations to improve overall sustainability of the NL CLARIN / CLARIAH services (Section 5).

2 Populating the NL CLARIN infrastructure

Activities for CLARIN were initiated in the Netherlands via the CLARIN-NL project and continued in the CLARIAH-NL projects.[5] A few projects were initiated centrally to implement basic infrastructural services, but the bulk of the services were user-

2 https://portal.clarin.nl

3 The terms CLARIN-NL and CLARIAH-NL refer to projects, which have created and extended the CLARIN and CLARIAH infrastructures in the Netherlands. For the latter we use the terms NL CLARIN and NL CLARIAH.

4 http://portal.clarin.nl/clariah-tools-fs

5 The CLARIAH-NL projects include the projects CLARIAH-SEED, CLARIAH-CORE and CLARIAH-PLUS.

driven and created in a series of four calls[6] over a period of five years (2011–2016). Invitations to submit proposals for projects for end-user facing services and tools, as well as infrastructural services for the benefit of the community, were issued and resulted in projects by small consortia of partners initially from the domain of Language Resources and Technology. This was followed up by the CLARIAH-NL projects, which have partially continued to support existing services but also added a number of new services to the NL CLARIN infrastructure.

In the original CLARIN-NL calls the strategy was explorative and expansive out of a desire to offer a broad set of organizations (university departments, research institutes, and general research support) the opportunity to get familiar with the initial CLARIN infrastructure components developed during the EU CLARIN preparatory phase by integrating their own data and services into CLARIN. An important reason for this explorative strategy was to investigate the needs of the broader humanities community: although CLARIN originated from the linguistics and computational linguistics communities, it aims to serve all humanities researchers working with language materials. At that time knowledge about the research questions and infrastructural needs of this broader class of humanities researchers was generally insufficient in the community that initiated CLARIN in the Netherlands.

CLARIN-NL tried to bring these two groups together so that humanities research questions could be shared and the potential of natural language processing and general infrastructural facilities for dealing with such research questions could be explored. This could then be translated into concrete plans for infrastructural facilities, and some of these were actually implemented.

As a consequence, many subprojects for CLARIN in the Netherlands were user-driven: we intentionally aimed for the selection of research topics, data, and supporting infrastructure facilities to be made by the researchers themselves. However, this resulted in many pieces of functionality that were highly tuned to a narrow class of specific research questions and often to a single corpus or dataset. We will provide several examples below, and characterize some of them in quite some detail. We do not hold their narrowness against these applications or the projects that developed them, because probably no one had the knowledge and expertise at the time to do it differently. And by encouraging applications from users we ensured a base interest in the topic. But now is a moment to reflect on this and to try to sketch of how they could be incorporated into more generic functionality.

6 http://www.clarin.nl/calls.html

3 Sustainability

The sustainability of services is not easy to ensure. Many factors play a role here, but we focus on the major ones that played a role in CLARIN in the Netherlands.

A first important factor is the organization that hosts the service. In the NL CLARIN context we always stipulated that only CLARIN B-centres should host services, though as will be shown below, we did not always succeed in enforcing this requirement. We also maintained the policy that only institutes with a longer-term mission to make data and services available for research purposes should become CLARIN B-centres in the Netherlands.[7] We discouraged research departments of universities from becoming CLARIN B-centres because their commitment to such a status is highly dependent on specific researchers or the specific research interests of one particular researcher, and therefore not sufficiently stable. Even if the researcher remains interested, there is no reason to expect commitment from the department or university to maintain the required infrastructural facilities (such as servers) for a longer period of time (Broeder et al. 2017). Of course, institutes with a longer-term mission to make data and services available are also not immune to changes and new developments. As shown below, we experienced our fair share of this in the Netherlands. But even then, such institutes are more stable than university research departments as service hosting centres.

A second factor is the degree to which a service is embedded in the hosting centre: if a service has been developed by the centre itself, or is actively used by the centre's employees, the commitment to keeping this service running is higher than for a service that has been developed by external developers or that has an external user base. As will be shown below, it happened regularly that a service developed by external developers and/or with a user base from outside the host had to be hosted by a centre, and this is generally not beneficial to its sustainability.

A third factor is the stability of the developer community. It will be easier to keep a service running if it has a solid and stable developer base. As will be shown below, this has often not been the case, even though measures were taken to improve the stability of the developer base.

Fourth, active use of a service by its targeted users, often leading to requests for new functionality or error reports, is generally beneficial for sustainability. It

7 Examples of such institutes in the Netherlands are the Meertens Institute, the Huygens Institute, the Institute for the Dutch Language, the Max Planck Institute for Psycholinguistics, and DANS.

keeps the maintenance of the service on the agenda and stimulates active search for funding the implementation of new functionality.

Finally, the number of services that must be maintained plays an important role in sustainability: in general, the smaller the number of applications, and the smaller the number of different components (frontend, backend) of such applications, the better it is for sustainability. Of course, a proper balance must be found here, because maintaining just a few extremely complex applications might also hinder sustainability. If one wants to achieve the same functionality with fewer applications, the applications have to be more generic in nature and cannot be too specific. Due to the setup of the initial CLARIN projects in the Netherlands, this has become a very important factor, as will be illustrated below via the case study into text search applications in the CLARIN infrastructure in the Netherlands.

3.1 Background

In order to understand the dynamics that underlie the variety of services, their institutional hosting and (challenges for their) sustainability, it is necessary to describe by which processes they came to be and are funded. Part of this background was already described in (Odijk and van Hessen 2017) and Section 2 "Populating the CLARIN NL Infrastructure".

Only a few technology requirements were imposed in the CLARIN-NL and CLARIAH-NL calls, in particular the requirements for interoperability within the larger CLARIN EU domain. Interoperability with CLARIN requires using CMDI[8] metadata (Broeder et al. 2010, 2011; Windhouwer and Goosen 2022) for describing resources, issuing Persistent Identifiers (PIDs) to identify resources, SAML-based Federated Identity Management (FIM) for authenticating users, and the use of a Server Oriented Architecture (SOA) to permit easy sharing of services by services.

A few of these interoperability requirements had to be relaxed for some partner organizations since they made different technology choices at an earlier stage. An example is the requirement to use the Handle System technology for PIDs, whereas DANS already used URN:NBN, and also waiving, or at least not enforcing, the requirement to use SAML-based FIM for allowing access to CLARIN services from outside of the Netherlands. That last requirement would sometimes require a change to the implemented accepted authentication option, which the service provider considered confusing for existing users. In addition, and

8 For an explanation of acronyms for technical components and standards, see Appendix 5.

especially for smaller software development groups, the required expertise for dealing with SAML-based FIM was lacking.

But these requirements contributed little to sustainability of the services, and no other requirements were imposed by CLARIN in the EU or in the Netherlands to ensure sustainability, in part because sustainability of services was largely uncharted territory. In this respect, we tried to learn from others who were ahead of us (inter alia via workshops with experts from the Software Sustainability Institute[9] and Knowledge Exchange),[10] but this started only as of 2013. However, it was difficult to see how adoption of these best-practices could be captured in requirements for the CLARIN-NL calls.

As stated, initially it was mostly organizations with a language research or language technology focus that responded to the calls, while later the response was broader also including other humanities disciplines and university libraries. The requirement that services must be hosted at a CLARIN B-centre was not only imposed for the stability and sustainability of the services and access to data, but also to foster the relationships of the CLARIN B-centres with their infrastructure specialists and research institutes with their humanities researchers. Unfortunately, we did not always succeed in having the services hosted by a CLARIN B-centre, especially for applications that were originally developed outside of CLARIN and highly interconnected with existing other parts of a research department's computational infrastructure. Examples of such services include PaQu,[11] WAHSP/BILAND,[12] TDS,[13] and WIP,[14] which will be discussed in more detail below.

3.2 Services classification

In this section we will discuss the major services, categorized into three classes: services targeting end users (Section 3.2.1), infrastructural services (Section 3.2.2), and services resulting from special collaborations (Section 3.2.3).

9 https://www.software.ac.uk/
10 https://www.knowledge-exchange.info/event/software-sustainability
11 https://portal.clarin.nl/node/14366
12 https://portal.clarin.nl/node/14383
13 https://portal.clarin.nl/node/14374
14 https://portal.clarin.nl/node/14386

3.2.1 Services and tools targeting end-users

Services and tools targeting end-users constitutes the largest group of services. Most are web applications that enable a user to search and browse through specific existing data-sets or corpora, and that have also a specific user interface for specifying queries and visualization. Most such services support only a fixed dataset, but some (e.g., PaQu, AutoSearch,[15] GRETEL 4[16]) allow the user to upload new data. Linguistic enrichment of new data is sometimes carried out by the search application (PaQu, GRETEL 4) but must be done with other services such as Frog,[17] TICCL,[18] or PICCL[19] outside the application. The resulting enriched data can then be uploaded in the search application (e.g., in AutoSearch). Such services may be essential for specific users and/or be broadly used, but they are not essential for the functioning of the infrastructure as a whole or even for other services, and will therefore not be missed if not used.

3.2.2 Infrastructural services

A second class consists of services that provide infrastructural services not directly seen by end-users. Many of these are currently provided by the CLARIN ERIC infrastructure and some strong B-centres that can afford to develop and host these. Such services require a strong commitment from the developing and hosting organizations in order to avoid long periods of minimal maintenance or even dysfunction,[20] since they are usually not immediately useful within the hosting organization, and receive less attention. Such services in the Netherlands are ISOcat,[21] CCR,[22] CLAVAS,[23] and CMD2RDF[24] (Windhouwer, Indarto, and Broeder 2017). These are basically registries, important for other services but not directly visible for end-users. Another class of infrastructural services are conver-

15 https://portal.clarin.nl/node/14324
16 https://portal.clarin.nl/node/14349
17 https://portal.clarin.nl/node/14344
18 https://portal.clarin.nl/node/1914
19 https://portal.clarin.nl/node/14392
20 Note that when it concerned infrastructure services essential for the operation of the EU wide CLARIN infrastructure, CLARIN ERIC took over their operation when dysfunctioning was imminent.
21 https://portal.clarin.nl/node/14353
22 https://portal.clarin.nl/node/14327
23 https://portal.clarin.nl/node/14330
24 https://portal.clarin.nl/node/14331

sion services such as Openconvert,[25] which also suffered from lack of resources for maintenance.

3.2.3 Special collaborations

Next to the regular calls, some of the services created by the CLARIN-NL projects were the results of projects with an emphasis on the collaborative aspect between partners, for example, TTNWW[26] (Kemps-Snijders et al. 2017), which was a collaboration between the Netherlands and Flanders. It produced a number of NLP workflows for both spoken and written text using existing NLP services. The collaboration aspect heavily influenced choices for architecture, which consisted of workflows of independently implemented NLP services provided as Virtual Machines (VMs), which were not anchored in the normal operations of the partners that provided these VMs. In addition, the VM hosting service provided by SURFsara for TTNWW was not guaranteed. It offered a good opportunity to learn and collaborate with this important Dutch academic IT service provider, but also caused frequent down-times aggravated by the need for specialized knowledge for restarting the TTNWW service.[27] Although this situation proved vulnerable with regard to sustainability of the TTNWW service as a whole (and currently the service is indeed unavailable), TTNWW met its main goals and under different circumstances might have evolved over time into a more stable and larger services framework. Other such special projects, from the CLARIAH-CORE project, are ATHENA,[28] and Amsterdam Time Machine.[29]

3.3 NL CLARIN services status in 2021

This section describes some relevant observations from our list of 85 services and tools that were created in the CLARIN-NL and CLARIAH-NL projects over a period of ten years. We base this on the CLAPOP[30] portal (Odijk 2019), where the results

25 https://portal.clarin.nl/node/14364
26 https://portal.clarin.nl/node/14378
27 Technologies such as docker-compose and Kubernetes, which were unavailable at that point, would have made a considerable difference.
28 https://clariah.nl/en/projects/athena-access-tool-historical-ecology-and-environmental-archeology
29 https://clariah.nl/en/projects/atm-amsterdam-time-machine
30 http://portal.clarin.nl/clariah-tools-fs

of the CLARIN-NL and CLARIAH-NL projects with regard to data provisioning and service building have been registered and from which the actual availability status was (manually) checked.[31] Some of the services listed in CLAPOP are general infrastructural services that are maintained in collaboration with and funded largely by CLARIN ERIC, such as ISOcat and its successor CCR. We exclude them from the sustainability discussion here since their maintenance and availability is steered from outside the NL CLARIN domain. Out of the 85 tracked services a small number must be considered lost, that is they are not on-line anymore and the originally responsible are no longer available or responding to enquiries. This is the case for seven of the listed services. For five other services it was made explicit that these were withdrawn, usually for reasons of technology obsolescence, e.g., Adobe Flash dependency for FESLI[32] and TDS, or dependence on specific environments, e.g., ANNEX, which depended on the obsolete LAT repository software (Kemps-Snijders et al. 2008). For four additional cases, the service was explicitly superseded by a new one, for example TiCClops[33] and COBWWWEB.[34] The manner in which end users are informed about service withdrawal or service succession varies by hosting organization, but almost no service description was complete without the hosting organization being specifically asked to update its service information pages. A large proportion of the tracked services (38) are web applications with functionality for searching in specific corpus content or databases. Some manage several such resources (e.g., the INT hosted dictionaries) but most are dedicated to one resource only. Two general engines were developed for searching through large corpora of linguistic information: MTAS (Brouwer, Brugman, and Kemps-Snijders 2016), and Blacklab (de Does, Niestadt, and Depuydt 2017). These are in use in end user facing services such as AutoSearch and OpenSoNaR[35] (Blacklab) and Nederlab[36] (MTAS). These also require considerable investment and expertise and are vulnerable when experts become unavailable, as happened in the case of MTAS. Although these general search engines would be prime candidates for technology merging, or for concentrating on the development of only one service, it proved very difficult to realize this because of aspects of partner institute autonomy and overlapping ambitions (see also Section 4). Only two services (registries) were true infrastructure services for the CLARIN infrastructure: CLAVAS and CCR. These are not intended for direct

31 This overview of the services will be replaced in 2022 by *ineo*.
32 https://portal.clarin.nl/node/14343
33 https://portal.clarin.nl/node/14376
34 https://portal.clarin.nl/node/14334
35 https://portal.clarin.nl/node/14365
36 https://portal.clarin.nl/node/14362

use by researchers and require special expertise to integrate them with other tools, which is how they should be used. CLAVAS proved not to be so essential since it was off-line for a long period without major problems. The CCR, however, is considered essential for central CLARIN operations and when Meertens was temporarily unable to support it, CLARIN ERIC took over.

3.4 Analysis

In this section we discuss four challenges for sustainability: reorganization of partner institutes that are CLARIN B-centres (Section 3.4.1), changing technologies (Section 3.4.2), the difficulty of maintaining the required expertise (Section 3.4.3), and service hosting (Section 3.4.4).

3.4.1 Reorganizing and restructuring of CLARIN centres

The reorganization and restructuring of partner institutes that were CLARIN B-centres did not only impact the sustainability of their services but rearranged the landscape with regard to the interest and capabilities of partners to continue their participation in the CLARIN commons. Over the past 10 years we have seen three major shifts in CLARIN B-centres in the Netherlands.

The first of these is a reorganization at the Institute for the Dutch Language (INT),[37] one of the NL CLARIN B-centres. For a long period it was unclear in which direction the institute would be heading. This created uncertainty for its employees but also about the role it could play in CLARIN. In the end, this reorganization did not have much impact on the availability of the services, nor on their further maintenance except for a period where the TST data[38] were unavailable. The INT ambitions and the available resources for this work have not changed since their initial participation in the CLARIN projects, which of course supports the sustainability of the services developed and hosted.

On the other hand, the changes at the MPI for Psycholinguistics (MPI-PL), which changed its ambitions in 2014 and decided to be involved only in infrastructure projects that directly were aligned with, and supportive of their immediate research interests, had a large impact. As the major CLARIN B-centre in NL,

37 At the time it was called the Institute for Dutch Lexicology (INL) and it also hosted the so-called 'TST-Centrale' (Language Technology Central).
38 https://ivdnt.org/taalmaterialen/

MPI-PL was very active in providing general infrastructure services (so-called Type A services) and it supported many external researchers. Although MPI-PL faithfully fulfilled its existing obligations, the necessary further software development and the hosting of services beyond direct MPI-PL interests were discontinued. For example, development support for tools such as ARBIL[39] for CMDI metadata editing and the LAT software stack, including a linguistic data repository stack, were terminated. Fortunately, CLARIAH-NL was able to move some services to other organizations and CLARIN ERIC took over responsibility for others. A positive side effect of the above is that, where the opportunity arose, new and better solutions were substituted for the old ones: CLARIN CCR for ISOcat (but with a different hosting organization), and the LAT software stack was replaced by the more modern Islandora-based FLAT repository system.

Thirdly, the clustering of three KNAW institutes (Meertens Institute, Huygens Institute, and the International Institute for Social History) into the Humanities Cluster (HuC), including two NL CLARIN B-centres, is the latest change to have a major impact on the CLARIAH services landscape. These institutes joined forces, inter alia to create a large pool of software developers to improve their working atmosphere, increase the possibilities of education, distribute their knowledge and expertise among multiple persons, and create career opportunities for the developers inside the HuC organization. Ironically enough, this did not prevent the two developers most knowledgeable about MTAS and some other services (TTNWW, PILNAR) from leaving during this reorganization process because they saw no viable future for them after this reorganization. Additionally, the reorganization efforts needed for integrating the three institutes' technical infrastructure (temporarily) took away resources for the planned support for and roll-out of new CLARIN services.

3.4.2 Changing technologies

Over a period of more than 10 years one would expect quite a few services and tools to be withdrawn or to become unusable because of their dependence on technologies no longer developed or having become inadequate, while the cost of upgrading to other technologies would be too steep. This was indeed clearly the case for some of the services depending on the Adobe Flash frontend (e.g., FESLI, PILNAR, TDS). It is notoriously difficult to make safe technology choices for graphical front ends. However, we also note that failing to update services

39 http://portal.clarin.nl/node/14320

with regard to advancing technology might also indicate a lack of interest from both providers and the project management, which should represent the end user and provide resource capacity. A more purposeful, coordinated way of dealing with obsolescence issues would be desirable and is perhaps feasible if more information on applied IT technologies and planned software updates can be tracked, for instance by adding such information separately to central service descriptions such as CLAPOP. Apart from changing technologies there is also the matter of advancing standards, which requires service updates. In our NL context we can think of CMDI as a metadata format and Folia as a linguistic data format. Fortunately the experts and developers involved with such updates are also often involved as implementers of tools using these standards. The tools mostly involved with CMDI, for example CMDI Forms for editing (Zeeman and Windhouwer 2018) and CMD2RDF for CMDI to RDF format conversion, are maintained at the Meertens Institute, which has CMDI experts who are also involved in CMDI standard advancement. With respect to updates of the Folia standard, some interoperability problems have been noticed that stem from insufficient coordination between the maintainers of different services using the Folia format. In situations where many different services depend on a common standard format, the process of updating common standards and adapting services should be coordinated properly, in order to prevent fragmentation in separate, non-interoperable islands.

3.4.3 Scarce expertise

In the CLARIN-NL and CLARIAH-NL projects, the project partners have had to manage challenges with regard to expert staff leaving, especially in times of reorganizations. This was certainly a cause for the withdrawal of some services, but also for the inability to repair or upgrade services when needed. The cost factor for producing academic software is such that it is very difficult to provide proper Service License Agreements (SLAs) and sufficient resources for maintenance and functionality enhancement in comparison with industry.

3.4.4 Service hosting

As already mentioned in the background section (Section 3.1), one of the requirements in the CLARIN-NL calls was the intention to host the resulting service (or data set) at one of the NL CLARIN B-centres, since these were considered to provide better service availability and sustainability. In some cases this led to

coincidental collaborations between the organizations responsible for service development and those doing the hosting. It also led to the hosting organization specifying extra requirements with regard to the service's expected environment and resource use, such as the use of a particular type of database or operating system version. This should be considered positive and it contributes to proper service operation and availability, but additional requirements imposed by the B-centres may also have motivated the software developers to (keep) hosting the services themselves. From the services listed on CLAPOP, ten are not hosted by CLARIN B-centres but for instance by university departments from Radboud University Nijmegen or from Groningen University. In addition, there are services hosted properly but outside of the direct CLARIN domain (e.g., at the National Institute for Sound and Vision, NISV). The WIP service, which is no longer available, was initially hosted by a development team at the University of Amsterdam, where the server hosting the service was discarded because it was considered obsolete, but not replaced. This is what one can expect from a research department that has no commitments for providing sustainable services, and this is why CLARIN B-centers, with a focus on sustainable access and stable services should be preferred. Nevertheless, many university departments have done an excellent job keeping services for which they have a specific long-term interest up-to-date and accessible for large groups of users. Therefore, we suggest that if there is no B-centre hosting candidate for a service, it is acceptable to have the service hosted by an organization that has an affinity with the service, even if that organization is not a B-centre. The CLARIN B-centres have not, overall, proven to be more stable than other organizations for services that were created in a small consortium consisting of a researcher and the CLARIN B-centre but that the centre was not interested in. The centres must also be more selective in accepting participation in such consortia.

4 Case study: Specificity and sustainability of search services

As was pointed out above, having a lot of different services is generally not beneficial for sustainability. In this Section we present a case study for one specific class of services: text search services. We argue that each of these services implements a different subset of the desired functionality, and that it is highly desirable to replace them with fewer, more generic services. This will improve sustainability but also the functionality for the user.

Apart from text search services, there are many other search services in CLARIN, but they will not be dealt with here systematically. Among these are services for searching in lexical resources, such as the historical dictionaries of Dutch and Frisian (ONW,[40] VMNW,[41] MNW,[42] WNT, and WFT-GTB[43]) in the historical dictionary portal[44] ANW,[45] DiaMaNT,[46] Cornetto,[47] Duelme,[48] GrNe,[49] and WebCelex.[50] There are also several services that enable search in structured data, for example for literary and historical data. Examples include Arthurian Fiction,[51] BNM-I,[52] COBWWWEB, DSS,[53] and Rembench.[54] There are also services for searching in structured linguistic data, such as TDS[55] and MIMORE[56] (Barbiers et al. 2016).

4.1 Specificity of search services

Many different text search applications have been developed in the CLARIN-NL and CLARIAH-NL projects in the Netherlands. In this section we will consider three subclasses: (1) applications for pure text search; (2) applications for search for text enriched with linguistic annotations at the token level; (3) applications for search in a treebank, that is, a text corpus in which each sentence has been assigned a syntactic structure.

40 http://portal.clarin.nl/node/14363
41 http://portal.clarin.nl/node/14381
42 http://portal.clarin.nl/node/14357
43 http://portal.clarin.nl/node/14385
44 https://gtb.ivdnt.org.
45 http://portal.clarin.nl/node/14319
46 https://diamant.ivdnt.org/diamant-ui/
47 http://portal.clarin.nl/node/14336
48 https://portal.clarin.nl/node/4200
49 http://portal.clarin.nl/node/14350
50 http://portal.clarin.nl/node/14384
51 https://portal.clarin.nl/node/4202
52 http://portal.clarin.nl/node/14326
53 https://portal.clarin.nl/node/4211
54 https://portal.clarin.nl/node/4227
55 https://portal.clarin.nl/node/14374
56 http://portal.clarin.nl/node/14356

4.2 Applications for pure text search

Search applications that are focused on searching purely for text (i.e. without any linguistic annotations) include PILNAR,[57] Polimedia,[58] WAHSP, BILAND,[59] TexCavator,[60] VK[61] and WIP from CLARIN-NL projects, and ePistolarium[62] from the CLARIN and CLARIN-NL supported but independently financed ePistolarium project (Ravenek, van den Heuvel, and Gerritsen 2017). The users who initiated these applications and use them are from humanities disciplines other than linguistics; they are therefore mostly interested in the content of the textual resource and have no specific interest in linguistic properties of these texts.

All applications offer the functionality to search for text using textual queries, often with support for Boolean operators. They also offer the option to narrow down the search to data meeting certain requirements on metadata. The metadata schema differs according to corpus. Most of these applications are highly specific and offer the ability to search in a single corpus – for instance, ePistolarium in correspondence between scholars in the 17th century in the Netherlands, PILNAR in a corpus of pilgrimage narratives, VK in the works of Lou de Jong on the Netherlands in World War II, and WIP in the proceedings of the Netherlands parliament.

Since these were different applications, developed independently of one another, it is not possible to carry out searches across multiple corpora, though that would obviously be useful in several cases. For example, the WIP project aimed to research mentions of World War II in the Dutch Parliament (WIP=War in Parliament), and a combined search in the parliamentary data and in the work of Lou de Jong on World War II as offered by VK would obviously be very useful. Polimedia did enable searching in multiple corpora, even corpora of different modalities: it links the minutes of the debates in the Dutch Parliament (Dutch Hansard) to the databases of historical newspapers and ANP radio bulletins to allow cross-media analysis of coverage in a uniform search interface through a combined search in these resources. WAHSP offered the ability to search in textual data from news media from the period 1863–1940 of the Dutch National Library. WAHSP was further developed into BILAND, which added the textual data from news media of the Staatsbibliothek zu Berlin, enabling bilingual

57 https://portal.clarin.nl/node/4214
58 http://portal.clarin.nl/node/14369
59 http://portal.clarin.nl/node/14383
60 http://portal.clarin.nl/node/14375
61 http://portal.clarin.nl/node/14379
62 http://portal.clarin.nl/node/14329

searching supported by a text translation service. Neither application runs any more, in part because no clear CLARIN-centre was identified for hosting the software, and in part because much of the software used was dependent on software only available on servers of the University of Amsterdam. In order to tackle these problems, the researcher involved had TexCavator developed and maintained by the NL eScience Centre, but it lacked most of the multilingual functionality of BILAND. On the other hand, it gave access to ShiCo (Shifting Concepts) (Martinez and Kenter 2018), developed independently by the NL eScience Center. ShiCo is a tool for visualizing concepts shifting over time, based on word2vec. Later still, the researcher involved transferred ShiCo's maintenance and further development to the Digital Humanities Lab of Utrecht University, which reimplemented it and has made it available as a new search application called iAnalyzer,[63] which offers search in multiple corpora; however, most of the advanced features have disappeared or are available for only a few of the corpora. Furthermore, in this application, one can search in only one corpus at a time. The corpora include several resources that have been licensed by Utrecht University from a commercial publisher and can currently only be used by employees of Utrecht University.

Summarizing, we observe the existence of many different search applications, each with their specific backend engine and own frontend, each developed by a different developer or development group. We also observe, on the one hand, that insufficient functionality is offered by each individual application (one can search only in a single corpus or a limited set of corpora at a time), while on the other hand, there is some duplication in functionality (the National Library newspaper archive can be searched through WAHSP and its successors and through Polimedia).

Many, but not all of the applications offer functionality that goes beyond the text-based search functionality. For example, BILAND offered sentiment mining, TexCavator analysis of shifts in concept over time through ShiCo, as well as some normalization, stemming, and stop word filtering. ePistolarium offers similarity search, and search using topic models. WIP offered the ability to search for text in combination with searching for and analysing metadata on the speaker (e.g., which party the speaker belongs to), which could also be nicely visualized. Many, but not all, offer various visualization options, e.g. word clouds, time lines, heat maps, and the like. But all this additional functionality is useful for all of these applications and for all of the corpora, so it would be much better if there were one generic application which includes all of this functionality for all corpora.

63 https://ianalyzer.hum.uu.nl

With so many different applications, different (small) developer teams and small user bases, it should come as no surprise that several of the applications do not run any more. For WAHSP, BILAND, and TexCavator this is to be expected and normal because they were replaced by iAnalyzer, though with significant loss of functionality and accessibility. For Polimedia it need not come as a surprise either, because its functionality has been integrated into the Media Suite[64] developed in the CLARIAH-CORE project, which is truly a development in the right direction. PILNAR does not run anymore because it used Flash software, which has become obsolete. The development team around PILNAR was small, and some of them left. It seems that the user community was also small and insufficiently influential, otherwise they would have instigated the hosting centre to keep the service running. The hosting institute lacked the means and, apparently, the inherent interest to replace the Flash software with an alternative to keep the service running, and the data have not been integrated in other search applications that are still running at the relevant institute. WIP was never hosted by a CLARIN B-centre, but by the developers at the University of Amsterdam, and does not run any more for the reasons described above.

4.3 Applications for search for linguistic annotations at the token level

Several search applications enable searches in text corpora in which linguistic annotations have been added to tokens ("token-annotated corpora"). These include AutoSearch, CHN',[65] COAVA,[66] Corpus Gysseling,[67] FESLI, NAMESCAPE,[68] Nederlab, OpenSoNaR, and SHEBANQ.[69] See Appendix A for an overview of their properties that are relevant in this context.

All of these search applications share the common functionality of being able to search for words, and word combinations, and, where available, grammatical properties of the tokens such as lemma, word form, part-of-speech tag, and inflectional information. All but COAVA and SHEBANQ use a query language based on the Corpus Query Processing (CQP) language (Evert and The OCWB Development Team 2010). This is, of course, good, but unfortunately each application sup-

64 https://mediasuite.clariah.nl/
65 https://portal.clarin.nl/node/14328
66 http://portal.clarin.nl/node/14333
67 http://portal.clarin.nl/node/14337
68 http://portal.clarin.nl/node/14358
69 https://portal.clarin.nl/node/4210

ports a different subset of CQP. Most allow filtering on the basis of metadata, but usually only before a search starts. Some applications share the same backend system (BlackLab, (de Does, Niestadt, and Depuydt 2017)), but each works with a different instantiation of this backend, thus complicating maintenance. Many also share the basic same front-end, but again each has a different instantiation and each differs from most if not all of the others. Several have options for analysing the search results. By "analysing search results" we mean, grouping, sorting, and/or filtering them, ideally in combination with metadata. This feature is, in our view, crucial for corpora with multiple annotations, especially since these annotations are not guaranteed to be 100% correct. The applications AutoSearch, CHN, and Corpus Gyseling all have more or less (but not exactly) the same system for analysis, which is limited, since one can generally analyse by a single criterion only (e.g., by part of speech, or by lemma, but not by these combined). Only OpenSoNaR allows analysis by multiple criteria, though not combinations of linguistic properties and metadata. One can, for example, create groupings of the data by grammatical properties, and see the relevant individual examples (or a subset thereof) by clicking on the grouping. Similarly, analysis of the search results in combination with metadata is possible but limited. Nederlab has even more limited options for analysing the search results: fewer options for grouping, no option to inspect the actual examples of a grouping. We do not know whether FESLI offered options for analysing the search results, and we can no longer check because it does not run any more, but we suspect that it did not offer this. NAMESCAPE and SHEBANQ do not offer any options for analysing the search results. COAVA enables the user to filter the search results by metadata and selecting nouns only.

As is obvious from this description, there are many different search applications for searches on token-annotated corpora, but each of them has limited options, a limited set of data that can be searched, and limited analysis options, and each implemented this in its own way. At the same time, there is also unnecessary duplication of functionality, for example for searching in the National Library news corpora archive. It is clear that with fewer and less varied applications more functionality can be added, the end user will need to learn less, and sustainability is increased.

There certainly are good developments as well here. As was pointed out above, many search applications are based on the BlackLab backend, and are based on the same basic frontend, and many are based on the same query language. Some search applications have functionality that would be useful in other search applications as well, for example the capability to store queries for reuse later and to share them with others is a helpful feature of SHEBANQ and Nederlab,

but this should be a feature for every search application.[70] Similarly, the feature of a combined search in a corpus and a lexicon, as offered by COAVA is functionality that would also be desirable for other search applications, for example to obtain properties of tokens from a search result in a lexicon such as CELEX or Cornetto[71] ("chaining search", (Dekker, Fanee, and de Does 2019; Odijk 2020)). The upload functionality offered by AutoSearch is very important, and it has been used quite extensively over the past five years, for a variety of projects, and also formed the basis for hosting Arabic corpora of Utrecht University developed in a collaboration project between the NL eScience Center and CLARIAH-NL.[72] The upload functionality also requires technology to automatically enrich a text corpus with linguistic annotations if one wants to search for linguistic properties. Such a pipeline was developed in the context of Nederlab, but the experts state that this pipeline is not suited for use by end users. However, one can use the Frog[73] (van den Bosch et al. 2007) web service via its web application interface, download the resulting data and upload them into AutoSearch. For languages other than Dutch one can use the pipelines defined in Weblicht,[74] and upload the results obtained from WebLicht into AutoSearch.[75,76]

The Nederlab project (Brugman et al. 2016), a project independent of CLARIN-NL and CLARIAH-NL but partially funded by them, was actually an attempt to create a single search application for the whole collection of Dutch historical textual data covering the period from 900–1900. This surely was a move in the right direction, because it would create a single search application for a huge amount of data. It was expected that the amount of data in which users could search would become so large that special measures were needed to ensure a reasonable performance of the system. There was close collaboration in the project between multiple partners, in particular Meertens Institute and the Institute for the Dutch Language (INT). INT had earlier developed the BlackLab search engine (de Does, Niestadt, and Depuydt 2017), which was in use for a lot of search applications, both for internal use and

70 The option of storing queries, however, also requires a way of organizing queries in such a way that they can be found back easily, and needs a user-specific store to store queries not shared with others.

71 http://portal.clarin.nl/node/14336

72 http://arabic-dh.hum.uu.nl/corpus-frontend/

73 https://webservices.cls.ru.nl/frog

74 https://weblicht.sfs.uni-tuebingen.de/weblichtwiki

75 It is certainly desirable to have such enrichment as part of the search application (as is possible in PaQu and GrETEL), at least as an option, because that makes enriching one's corpus much easier for the user.

76 See https://surfdrive.surf.nl/files/index.php/s/JkYKlHSNZnj7ysJ for a recorded lecture, a presentation and relevant materials to illustrate this.

for use by external researchers. Meertens did not have a search engine. It would have been natural to start from BlackLab and modify and extend it so that it could deal with the expected volume of data. However, for reasons of autonomy and efficiency, Meertens Institute, which was leading the project, decided to develop a completely new backend from scratch (the MTAS-engine: Multi Tier Annotation Search, (Brouwer, Brugman, and Kemps-Snijders 2016)). This was a risk of course, but defensible since Meertens also has the obligation to build up knowledge and expertise in providing search applications for research purposes. An additional problem, however, was that the MTAS development team was rather small: in essence, two people. As described above, these very two developers left during this reorganization process intended to strengthen sustainability. As a consequence, only limited knowledge of and expertise with MTAS is available now, and we must see how this will develop in the near future. Hopefully, some consolidation of the Blacklab and MTAS efforts can take place.

4.4 Applications for search in treebanks

A treebank is a text corpus in which each sentence has been assigned a syntactic structure. Syntactic structures are often trees, hence the name 'treebank' for such corpora. Examples of applications for search in treebanks are Lassy Search,[77] PaQu, GRETEL 1-4, and Corpus Studio Web.[78]

Lassy Search was originally developed outside of CLARIN-NL though clearly inspired by the desire expressed by CLARIN to make corpus searching easier for non-expert users. It offered the ability to search for grammatical relations between two words in the Lassy-Small Corpus, via a dedicated interface.

This application was not systematically maintained, and when a need for additional functionality arose, a new version, called PaQu, was developed. PaQu offers the ability to search not only for grammatical relations between words via a dedicated interface, but also via Xpath queries. It enables users to search in additional corpora (initially only the Spoken Dutch Corpus, currently several more), and enables a user to upload his/her own corpus. This corpus is automatically parsed by Alpino and the resulting treebank is made available for searching. PaQu also extended the options for (limited) analysis of the search results, and

77 http://www.let.rug.nl/~alfa/lassy/bin/lassy-save
78 http://portal.clarin.nl/node/14338

allows macros to simplify queries and make queries or parts of them reusable (Odijk et al. 2017).[79]

GRETEL (Augustinus et al. 2017) was originally developed by KU Leuven in the context of the cooperation between the Netherlands and Flanders on CLARIN. It originally offered search in the Lassy-Small Corpus and the Spoken Dutch Corpus. Its distinguishing feature is the query by example option: the user can enter an example sentence that illustrates the construction they are interested in and select via a dedicated interface which aspects of this example sentence are crucial for the construction. After that an Xpath query is automatically created by the system and a search is started in the desired corpus. GRETEL also offers the ability to search with Xpath queries.

GRETEL 4 (Odijk, van der Klis, and Spoel 2018) extended the original GRETEL application (which had already gone through three different improved versions) and added two major new functionalities: (1) the option to upload one's own corpus (similar functionality as described for PaQu above), and (2) extensive options for analysing search results in terms of properties of the nodes that match with node descriptions in the Xpath query, in combination with metadata. A user can compose a pivot table in a graphical interface by selecting node properties and metadata in arbitrary combinations of indefinite size and drag them to the table.

Corpus Studio Web (Komen 2017) enables search in treebanks using XQuery and offers a query wizard to make the creation of queries easier. It has a completely independent origin, offers yet another mode of search in treebanks and includes more functionality than search alone.

It is obvious that PaQu and GRETEL 4 have large overlap in terms of the provided functionality. The types of corpora that can be offered for search are similar (and largely overlapping), both offer XPath search, both offer the service for users to upload their own corpora. The crucial difference between the two applications is the dedicated search options they offer: word relation search in PaQu and query by example in GRETEL. But the systems have been implemented differently (e.g., they use different XML-database systems, the programming languages used differ), which also leads to differences in the kind of Xpath queries one can formulate, and there are other differences as well: for example, the options for analysing search results are more limited in PaQu. It is obvious that it would be much preferable to have a single application combining the two distinguishing user interfaces in one application, combining all the corpora offered by

79 PaQu also formed the basis for the SPOD application (van Noord et al. 2020; Hoeksema, de Glopper, and van Noord 2022), but we leave this aside here.

the separate applications, using the best database engine for these systems after an evaluation of the available options, and the search result analysis options of GRETEL because they are more powerful than those of PaQu, the sample selection methods provided by PaQu (but not by GRETEL), the macro options of PaQu since they are better than the ones offered by GRETEL, and so on. There is a long wish list of additional functionality in these applications, which then has to be implemented only once. And it makes sense to investigate whether Corpus Studio Web can be involved in such an integration as well.

The PaQu and GRETEL applications were developed with linguistic research as main intended use. But the syntactic analyses that they offer might be useful for disambiguation purposes in other contexts as well. It is therefore desirable to integrate the treebank search and analysis options in a more generic search application that also offers pure text searching and the ability to search for token-based annotations.

4.5 Sustainability of search services

Since such a large proportion of the NL CLARIN services are in essence specialized search services optimized for specific structured information or data, it should be useful to analyse their existence and evolution in more detail.

As we have seen above, each search application in the NL part of the CLARIN infrastructure offers a different subset of the desired functionality, and each has data- and research goal-specific extensions that are actually useful for other data as well. Each application has its own frontend and backend. In short, we see a highly fragmented situation, which is difficult to maintain over a longer period of time. It is therefore desirable to reduce the number of different applications, backends, and front-ends, and to offer the union of the different functionality subsets in the (reduced number of) applications. This will increase the functionality for the user and increase sustainability.

One might be tempted to suggest that there should be a single instantiation of a single search application in the whole CLARIN infrastructure. That would optimize the prospects for sustainability. However, this is not feasible, for several reasons. First, a single instantiation and a single application imply a single point of failure, so it reduces robustness, which is also a desirable feature of infrastructural facilities. Second, it is not obvious how large the developer community could be, and what the commitment of the individual developers to a central system would be. Third, and most important: the data that are to be searched in are distributed over multiple centres in multiple countries. It is not desirable and

not feasible (for legal and technical reasons) to bring all these data together in a central place where the search application runs.

One might consider the option of having one search application per CLARIN member, but this is not in general desirable or feasible. A more natural approach is to have one search application per CLARIN B-centre that makes data available for users to search. After all, most centres want and have the obligation to build up knowledge and expertise to provide data and the capability to search within the data to their clients (researchers). Most CLARIN B-centres are also research institutes, and they offer data and the capability to search within the data to enable their researchers to carry out the institute's research goals. Ideally, each centre combines its obligations to its own researchers and research purposes with the CLARIN requirements. With just a single search application in each institute, the possibilities to reduce the dependence on a single developer or a very small number of developers can be more easily reduced, though this also requires a certain scale (the developing team of the institute must not be too small) and an intentional institute policy to spread the knowledge and expertise among its developers so as to reduce this dependence.

We recommend that CLARIN initiates a description of the desired functionality of a local search application that supports keyword search, lexical and grammatical search and mixed corpus and lexicon search for specific corpora but also for new corpora that a user can submit to the service, supported by linguistic and other enrichment pipelines (POS-tagging, parsing, named entity detection and linking, language detection, etc.), as well as offering a framework for plugging in new advanced services such as topic detection, word-embedding based search, facilities to deal with multilingual corpora, linking to external knowledge sources, etc. The description of the desired functionality must, of course, be regularly updated to reflect new developments. In such a more generic search application, covering multiple corpora, one should keep the metadata associated to the different corpora separate, at least in the first stage of integration. At a later stage one can start integrating the metadata. Of course, there will always be metadata properties that are unique to a corpus, but many of them are shared among all or a significant class of resources. For example, resource properties such as title, publication date, publisher, OCR-confidence, and author properties such as author name, author age, author birthday, author place of birth, author death date, author place of death, and author gender recur in many resources and can probably be relatively easily harmonized. The property genre or category also often recurs, but may be more difficult to harmonize. The search functionality will increase in power to the extent that these metadata have been harmonized.

It should also be clearly defined which data formats and other standards (e.g. for semantic operability) are supported by this search application. Obviously, it

should cover most data formats that are actually in use, but a small set might be particularly preferred. Applications such as AutoSearch, PaQu and GRETEL currently already provide such a list of supported formats. Any researcher or data provider can include his/her own data simply by ensuring that it is in one of the supported formats.

With a single application covering a large collection of data, there is of course the danger that a user who is interested in only a single dataset will suffer from the presence of this large collection (most of which he/she is not interested in). It should therefore be easy for a user to restrict search to a subset of the full collection, and to store the selection option so that this option is automatically selected in each next session until the user decides to modify it.

A single application that offers multiple search modes (such as e.g. the simple, extended, advanced, and expert modes of OpenSoNaR) must also ensure that there are multiple interface options, which can be selected depending on the expertise of the user and the character and complexity of the search query.

More generally, it requires careful investigation in each case as to whether search options in a dataset should be offered in a search application that also cover other datasets and/or other search options, or in a separate dedicated application, but for the situation in the Netherlands as sketched above the conclusion is obvious to us. Of course, with one search application per CLARIN B-centre, it is not possible to search across data that resides on servers of different centres. Federated content search (FCS)[80] (Stehouwer, Ďurčo, and Broeder 2012) should make that possible. CLARIN, of course, already worked on FCS, initially for pure text search, at a later stage also for search in token-annotated corpora. But the functionality of FCS should be extended to cover all the options that local search offers, which includes text search, search for grammatical properties, search in treebanks, search for metadata, analysing (grouping, sorting, filtering) search results in combination with metadata, and so forth, and not just the intersection of what all local search applications offer. FCS requires that a FCS endpoint is created for each local search backend and this requires a detailed specification of the character and format of the queries the endpoint must be able to process, and of the character and format of the search and analysis results that it returns to the FCS aggregator. The FCS frontend should offer all the functionality that the frontends of the local search applications offer. The work on developing this specification and its implementation, which has already been started by CLARIN, should therefore be continued, and it may also serve in part as a specification of the functionality that the local search applications should offer. It should be a

80 See https://www.clarin.eu/content/federated-content-search-clarin-fcs

CLARIN policy to commit many central resources to this topic, and to stimulate (or even require) CLARIN members to contribute to FCS via their national projects.

5 General recommendations for improving service sustainability

From our observations and background knowledge on ten years CLARIN service development and funding, we are able to make some recommendations:

1. The need for adequate reliable tracking of service hosting and maintenance history and performance, in addition to public relations and outreach effort and means to measure service uptake in specific domains and organizations: analysing papers and citations, measuring clicks, etc.
2. Such a service registry could be used also for dealing with software obsolescence issues in a coordinated way, maintaining information with regard to applied IT technologies and planned software updates can be helpful to predict and plan for necessary upgrades from a central project level.
3. A service hosting organization should host services that fall within its scope, i.e., align with its own mission and research goals. This is preferably a certified CLARIN B-centre, but it is more important that the hosting organization conforms to interoperability requirements such as, for instance, SAML-based authentication for AAI. Note that technology advancements such as containers make it relatively easy in the case of scalability or computing resource issues to host such services at general academic or commercial hosting providers.
4. Since, compared with the start of the CLARIN-NL project, we now have a sufficiently large consortium of relevant partners involved with creating and using research infrastructure, funding can be more specifically targeted at sustainability aspects, such as making the services part of their own internal research work flows.
5. For selected tasks and application types, specific policies should be agreed to increase efficiency and sustainability:
   (a) For example, for searching in token-annotated corpora there should be as few different search applications as possible,preferably at most one per CLARIN B-centre.
   (b) CLARIN should initiate a description of the desired functionality of a local search application that supports keyword search, lexical and grammatical search, and mixed corpus and lexicon search for specific corpora but also for new corpora that a user can submit to the service (supported

by linguistic and other enrichment pipelines (POS-tagging, parsing, named entity detection and linking, language detection, etc., etc.), as well as offering a framework for plugging in new advanced services such as topic detection, word-embedding based search, facilities to deal with multilingual corpora, linking to external knowledge sources, etc.

Appendix A: Token-Annotated Search applications

App	Data	3	4	5	Backend	Dedicated Interfaces	Search Result Analysis	Languages
AutoSearch	a user's own data	+	+	CQP subset	BlackLab	4	yes	Language independent
CHN	Contemporary Dutch Corpus	+	+	CQP subset	BlackLab	2	yes	Dutch
COAVA	CHILDES	+	–	none	idiosyncratic	yes	no	Dutch
Corpus Gysseling	Corpus Gysseling	+	+	CQP subset	BlackLab	4	yes	13th Century Dutch
FESLI	BISLI CHAT data	+	+	CQP subset	idiosyncratic	no	no	Dutch
Namescape	novels	+	–	none	idiosyncratic	yes	no	Dutch
Nederlab	Dutch texts 900–1900	+	+	CQP subset	MTAS	3	limited	Different historical variants of Dutch
OpenSoNaR	Contemporary Written Dutch Corpus	+	+	CQP subset	BlackLab	4	yes	Dutch
SHEBANQ	Bible texts	+	+	MQL	EMDROS	no	no	Hebrew, Syriac

Column 4 specifies *string search*, column 5 *token search*, and column 5 query language.

Appendix B: Acronyms

Acronym	Expansion	Clarification	URL
CMDI	Component Metadata Infrastructure	Metadata infrastructure required by CLARIN	https://www.clarin.eu/content/component-metadata
FCS	Federated Content Search	Distributed text search infrastructure promoted by CLARIN	https://www.clarin.eu/content/federated-content-search-clarin-fcs
FIM	Federated Identity Management	CLARIN requires SAML based FIM	https://en.wikipedia.org/wiki/Federated_identity#Management
SOA	Server Oriented Architecture		https://en.wikipedia.org/wiki/Service-oriented_architecture
PID	Persistent Identifier		https://en.wikipedia.org/wiki/Persistent_identifier
URN:NBN	Universal Resource Identifier/National Bibliography Number	Publication Identifier system	https://www.ifla.org/files/assets/bibliography/national_bibliography_number.pdf
HS	Handle System	PID technology promoted and required by CLARIN	https://en.wikipedia.org/wiki/Handle_System
SAML	Security Assertion Markup Language	A technology enabling Federated Identity Management and Single Sign-On authentication	https://en.wikipedia.org/wiki/Security_Assertion_Markup_Language
VM	Virtual Machine		https://en.wikipedia.org/wiki/Virtual_machine

Bibliography

Augustinus, Liesbeth, Vincent Vandeghinste, Ineke Schuurman & Frank Van Eynde. 2017. GrETEL: A tool for example-based treebank mining. In Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the low countries*, 269–280. London, UK: Ubiquity. DOI: http://dx.doi.org/10.5334/bbi.22. License: CC-BY 4.0.

Barbiers, Sjef, Marjo van Koppen, Hans Bennis & Norbert Corver. 2016. MIcrocomparative MOrphosyntactic REsearch (MIMORE): Mapping partial grammars of Flemish, Brabantish and Dutch. *Lingua* 178: 5 – 31. Linguistic Research in the CLARIN Infrastructure.

Bosch, Antal van den, G.J. Busser, Walter Daelemans & S. Canisius. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. In Frank Van Eynde, Peter Dirix, Ineke Schuurman & Vincent Vandeghinste (eds.), *Selected papers of the 17th computational linguistics in the Netherlands meeting*, 99 – 114. Leuven, Belgium: KU Leuven.

Broeder, D., M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg & C. Zinn. 2010. A data category registry- and component-based metadata framework. In N. Calzolari, B. Maegaard, J. Mariani, J. Odijk, K. Choukri, S. Piperidis, M. Rosner & D. Tapias (eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC 2010)*, 43–47. Valetta, Malta: European Language Resources Association (ELRA).

Broeder, Daan, Jan Theo Bakker, Marco van der Laan, Marc Kemps-Snijders, Menzo Windhouwer & Marjan Grootveld. 2017. Building CLARIN infrastructure in the Netherlands. In Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the low countries*, 45–59. London, UK: Ubiquity. DOI: http://dx.doi.org/10.5334/bbi.24. License: CC-BY 4.0.

Broeder, Daan, Oliver Schonefeld, Thorsten Trippel, Dieter Van Uytvanck & Andreas Witt. 2011. A pragmatic approach to XML interoperability – the Component Metadata Infrastructure (CMDI). *Proceedings of balisage: The markup conference 2011*. https://www.balisage.net/Proceedings/vol7/print/Broeder01/BalisageVol7-Broeder01.html.

Brouwer, Matthijs, Hennie Brugman & Marc Kemps-Snijders. 2016. A Solr/Lucene based multi tier annotation search solution. *Selected papers from the CLARIN annual conference 2016, 26–28 October, Aix-en-Provence*, 29–37. Linköping, Sweden: Linköping University Electronic Press.

Brugman, Hennie, Martin Reynaert, Nicoline van der Sijs, René van Stipriaan, Erik Tjong Kim Sang & Antal van den Bosch. 2016. Nederlab: Towards a single portal and research environment for diachronic Dutch text corpora. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA).

Dekker, Peter, Mathieu Fanee & Jesse de Does. 2019. CLARIAH chaining search: A platform for combined exploitation of multiple linguistic resources. In K. Simov & M. Eskevich (eds.), *Proceedings of CLARIN annual conference 2019*, Theory and Applications of Natural Language Processing, 24–27. CLARIN.

Does, J. de, J. Niestadt & K. Depuydt. 2017. Creating research environments with BlackLab. In Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the low countries*, 245–257. London, UK: Ubiquity. DOI: http://dx.doi.org/10.5334/bbi.20. License: CC-BY 4.0.

Evert, Stefan & The OCWB Development Team. 2010. The IMS Open Corpus Workbench (CWB): CQP Query Language Tutorial. OCWB report, IMS, Stuttgart. http://cwb.sourceforge.net/files/CQP_Tutorial/.

Hoeksema, Jack, Kees de Glopper & Gertjan van Noord. 2022. Syntactic profiles in secondary school writing using PaQu and SPOD. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources.* Berlin: De Gruyter.

Kemps-Snijders, Marc, Alex Klassmann, Claus Zinn, Peter Berck, Albert Russel & Peter Wittenburg. 2008. Exploring and enriching a language resource archive via the web. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis & Daniel Tapias (eds.), *Proceedings of the sixth international conference on language resources and evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Kemps-Snijders, Marc, Ineke Schuurman, Walter Daelemans, Kris Demuynck, Brecht Desplanques, Véronique Hoste, Marijn Huybregts, Jean-Paul Martens, Hans Paulussen, Joris Pelemans, Martin Reynaert, Vincent Vandeghinste. Antal van den Bosch,Henk van

den Heuvel,Maarten van Gompel,Gertjan van Noord &Patrick Wambacq 2017. TTNWW to the rescue: No need to know how to handle tools and resources. In Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the low countries*, 83–93. London, UK: Ubiquity. DOI: http://dx.doi.org/10.5334/bbi.7. License: CC-BY 4.0.

Komen, Erwin. 2017. Beyond counting syntactic hits. In Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the low countries*, 259–268. London, UK: Ubiquity. DOI: http://dx.doi.org/10.5334/bbi.21. License: CC-BY 4.0.

Martinez, Carlos & Tom Kenter. 2018. ShiCo – Exploring Shifting Concepts Through Time. DOI: 10.5281/zenodo.1435021.

Noord, Gertjan van, Jack Hoeksema, Peter Kleiweg & Gosse Bouma. 2020. SPOD: Syntactic profiler of Dutch. *Computational Linguistics in the Netherlands Journal* 10 (Dec.): 129–145.

Odijk, Jan. 2019. Discovering software resources in CLARIN. *Selected papers from the CLARIN annual conference 2018, Pisa, 8–10 October 2018*, Linköping Electronic Conference Proceedings no. 159, 121–132. Linköping University Electronic Press, Linköpings universitet. https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=159 &Article_No=13.

Odijk, Jan. 2020. De verleidingen en gevaren van GrETEL. *Nederlandse Taalkunde* 25 (1): 7–38.

Odijk, Jan & Arjan van Hessen. (eds.) 2017. *CLARIN in the low countries*. London, UK: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bbi. License: CC-BY 4.0.

Odijk, Jan, Martijn van der Klis & Sheean Spoel. 2018. Extensions to the GrETEL treebank query application. *Proceedings of the 16th international workshop on treebanks and linguistic theories (tlt16)*, 46–55. Prague, Czech Republic. http://aclweb.org/anthology/W/W17/W17-7608.pdf.

Odijk, Jan, Gertjan van Noord, Peter Kleiweg & Erik Tjong Kim Sang. 2017. The parse and query (PaQu) application. In Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the low countries*, 281–297. London, UK: Ubiquity. DOI: http://dx.doi.org/10.5334/bbi.23. License: CC-BY 4.0.

Ravenek, Walter, Charles van den Heuvel & Guido Gerritsen. 2017. The ePistolarium: Origins and Techniques. In Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the low countries*, 317–323. London, UK: Ubiquity. DOI: http://dx.doi.org/10.5334/bbi.26. License: CC-BY 4.0.

Stehouwer, Herman, Matej Ďurčo & Daan Broeder. 2012. Federated search: Towards a common search infrastructure. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Windhouwer, Menzo & Twan Goosen. 2022. Component Metadata Infrastructure. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Windhouwer, Menzo, Eko Indarto & Daan Broeder. 2017. CMD2RDF: Building a bridge from CLARIN to Linked Open Data. In Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the low countries*, 95–103. London, UK: Ubiquity. DOI: http://dx.doi.org/10.5334/bbi.8. License: CC-BY 4.0.

Zeeman, Rob & Menzo Windhouwer. 2018. Tweak your CMDI Forms to the Max. In Inguna Skadiņa & Maria Eskevich (eds.), *Proceedings of the 2018 CLARIN annual conference*, 95–98. Pisa, Italy. https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf.