



# The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond

A. Branco<sup>1</sup> · M. Eskevich<sup>2</sup> · F. Frontini<sup>3</sup> · J. Hajič<sup>4</sup> · E. Hinrichs<sup>5</sup> · F. de Jong<sup>2</sup> · P. Kamocki<sup>6</sup> · A. König<sup>2</sup> · K. Lindén<sup>7</sup> · C. Navarretta<sup>8</sup> · M. Piasecki<sup>9</sup> · S. Piperidis<sup>10</sup> · O. Pitkänen<sup>11</sup> · K. Simov<sup>12</sup> · I. Skadiņa<sup>13</sup> · T. Trippel<sup>14</sup> · A. Witt<sup>15</sup> · C. Zinn<sup>5</sup>

Accepted: 28 March 2023  
© The Author(s) 2023

## Abstract

CLARIN is a European Research Infrastructure Consortium developing and providing a federated and interoperable platform to support scientists in the field of the Social Sciences and Humanities in carrying-out language-related research. This contribution provides an overview of the entire infrastructure with a particular focus on tool interoperability, ease of access to research data, tools and services, the importance of sharing knowledge within and across (national) communities, and community building. By taking into account FAIR principles from the very beginning, CLARIN succeeded in becoming a successful example of a research infrastructure that is actively used by its members. The benefits CLARIN members reap from their infrastructure secure a future for their common good that is both sustainable and attractive to partners beyond the original target groups.

**Keywords** Research infrastructure · Interoperability · Social sciences and humanities · Language resources · Language technology

## 1 Introduction

Since its inception, CLARIN (see Fišer & Witt, 2022) has operated in line with the European agenda for Open Science, and it can be seen as an early adopter of the FAIR data principles (Collins et al., 2018; Wilkinson et al., 2016) *avant la lettre*<sup>1</sup> (Fišer et al., 2018). For CLARIN, the adoption of the FAIR principles was not a simple design decision, to be taken among many others, but the central one, which was forced upon CLARIN by its distributed nature. While starting with a core group of national consortia, by November 2022, 25 countries and over 50 centres

---

<sup>1</sup> <https://www.clarin.eu/fair>.

were involved in CLARIN.<sup>2</sup> The consortia represent research on language-related resources from a wide range of different scientific angles, make use of a diverse set of scientific methods, harbour large quantities of multi-flavoured research data, and give access to a broad set of processing tools. The geographically distributed nature of the CLARIN community helped conceiving the functioning of the CLARIN infrastructure as a federation of services, which in turn is driven by a number of central services that help users to discover, access and use research data and tools.

The building blocks of the CLARIN federation are the individual CLARIN centres for which three types of centres have been defined: the Service Providing Centres that establish the technical backbone of CLARIN, Metadata Providing Centres that feed metadata into central places, and Knowledge Centres that ensure that knowledge and expertise is shared within and across national boundaries.<sup>3</sup> Centres have to conform to technical and organisational principles and standards to ensure a seamless integration between the distributed resources available in the centres and the available services across the federation. All centres are hence certified to ensure that such conditions are met.<sup>4</sup>

From the outside, the success of CLARIN is clearly visible by its increasing number of national consortia and the many centres contributing to the overall CLARIN community. From the inside, CLARIN is best described as a well-oiled machine whose performance is driven by interoperability taken seriously.

CLARIN, hence, was well positioned to face a range of additional interoperability requirements on both the technical and organisational levels when it participated in the emerging European Open Science Cloud (EOSC), envisaged to be an open and trusted environment for managing data from all research domains. EOSC will federate the service offers from both existing and emerging data infrastructures, and is meant to become the universal access channel to all registered data repositories and cloud-based services through which all European researchers will be able to access, use, and reuse research outputs and data across disciplines.<sup>5</sup>

Preparatory to EOSC integration has been CLARIN's participation in the H2020 project *SSH Open Cloud* (SSHOC) within the Social Sciences & Humanities (SSH) cluster. Several of CLARIN's central services (such as the Virtual Language Observatory, the Language Resource Switchboard and the Virtual Collection Registry; see section 2), as well as the available knowledge and expertise are now being opened up for use beyond our traditional communities of reference. This is contributing to the uptake of shared services and practices across communities and disciplines. From the point of view of user involvement, the participation in a series of EOSC-related projects has offered CLARIN members the opportunity to exchange expertise and practices with partners from new domains. Examples include joint work with regard to legal issues, and in particular the rights of data

<sup>2</sup> There are 22 full members, 2 countries having *observer* status, and 1 third-party member, see <https://www.clarin.eu/content/participating-consortia>.

<sup>3</sup> <https://www.clarin.eu/content/clarin-centres>.

<sup>4</sup> For instance, a CLARIN centre that provides a repository service will need to apply for the CoreTrust-Seal (<http://coretrustseal.org>).

<sup>5</sup> <https://www.eosc-portal.eu>.

subjects, but also the provision of training sessions on, say, oral and written corpora targeted at social scientists and librarians.

CLARIN must take into account new developments in science that often have an impact on the interoperability aspect of its infrastructure. Our aim is to initiate and support the application of novel, data-driven methods for the SSH domain, which has been enabled by the developments in Digital Humanities and Computational Social Sciences, and by treating language data as a rich resource for disciplinary fields such as Data Science, Language Technology, and Artificial Intelligence. A careful reader will notice that the multitude of projects and perspectives discussed in this paper perfectly reflects the heterogeneity of an internationally distributed research infrastructure, catering to a wide range of needs within the academic community and beyond.

The remainder of the paper has three main parts. In Sect. 2, we describe first and foremost the technical infrastructure of CLARIN and how we succeeded to build a central common gateway to language-related research data, tools, and services hosted in a distributed environment. A well-oiled technical infrastructure must be complemented by a knowledge-based infrastructure that ensures that users know about the technology and how it can be effectively used to support their research. In Sect. 3, we describe the CLARIN Knowledge Centres and the important role they play in the community for sharing knowledge and expertise. The community building aspect is covered in a dedicated section, Sect. 4. Community building is important and a continuing task to ensure that CLARIN is thriving by onboarding new users and their needs, which in turn feeds into technical requirements and which enriches the knowledge infrastructure for all to profit. It is well worth pointing out that industry takes notice, and that CLARIN technology is started to get used beyond academic research in the Social Sciences & Humanities.

## 2 CLARIN as a provider of tools, data and services

The CLARIN community shares a long history when it comes to the collection, manual annotation and automatic processing of linguistic data. With an ever increasing number of data sets and tools available, it became necessary to develop a common metadata framework for the description of such linguistic data, and to devise technological solutions so that resources can be found and tools invoked with ease. The WebLicht workflow engine and the Language Resource Switchboard are two central pillars in CLARIN's technical infrastructure that guide users to find and invoke tools that can process their data (see Sect. 2.1). The Virtual Language Observatory is a faceted browsing environment empowering users to explore the CLARIN linguistic data space of over a million resources via a combination of faceted and full-text search. The use of a common metadata framework greatly facilitates the description of linguistic data and their organisation into faceted sub-spaces (Sect. 2.2). Giving users access to data and tools, where licensing issues have to be taken into account, is facilitated by a common authentication and authorisation infrastructure (Sect. 2.3). The reproducibility of scientific results is central to the credibility of the scientific endeavour. By leveraging its unique resources in terms of

expertise and technological assets, CLARIN is pioneering efforts towards the reproducibility challenge in the area of language science and technology (Sect. 2.4).

## 2.1 Weblicht, Switchboard, LAPPS grid

Over the years, the CLARIN community has developed a wide range of tools to process language-related data. Tool discovery and interoperability soon became an issue that CLARIN addressed in terms of a centrally available language technology platform giving users easy access to geographically distributed tools. WebLicht and the CLARIN Language Resource Switchboard are integral parts of the platform.

### 2.1.1 The WebLicht workflow engine

WebLicht (Hinrichs et al., 2010) is an environment for building and executing pipelines of natural language processing tools, with integrated capabilities for visualising and searching the resulting annotations.<sup>6</sup> The primary goal of WebLicht is to provide easy access to a wide range of text annotation tools to researchers in the Humanities and Social Sciences. WebLicht's annotation tools can be invoked via any web browser, without the need for local software installation or any prior familiarity with the tools.

WebLicht is built upon Service Oriented Architecture (SOA) principles: the processing tools WebLicht gives access to are implemented as web services across the web. The main components of WebLicht and their interactions are shown in Fig. 1. WebLicht is tightly integrated into the CLARIN infrastructure (Dima et al., 2012) and utilizes key components of it, such as the CLARIN Centre Registry and the CLARIN Identity Federation (for the latter, see Sect. 2.3).

Using metadata information from the CLARIN Centre Registry, the harvester polls each centre's repository for WebLicht-compatible web service metadata at regular time intervals. The pipelining and execution engine uses the harvested web service description metadata and the profile of the input data to create valid workflows for adding linguistic annotations (e.g., part-of-speech tags, lemmas, morphology) to text. The engine performs three important tasks:

1. It determines which services can be added to the pipeline. For example, if one of the tools in the pipeline produces lemma information, other services that produce lemmas will be omitted from the list. Similarly, in a pipeline that does not produce lemmas, any services that require lemmas in the input will not appear in the list. The resulting list is presented to the user for selection. This process is repeated each time the user adds or removes a service from their pipeline.
2. It ensures that a pipeline is valid.
3. It executes the pipeline, which is carried out by sequentially sending HTTP POST requests to the services, where the body of the request to service  $n + 1$  is the response of service  $n$ .

<sup>6</sup> <https://weblicht.sfs.uni-tuebingen.de>.

To address the difficulties arising from the fact that each tool has its own input and output formats, TCF (Text Corpus Format, Heid et al., 2010) was developed for use as an internal data exchange format. Annotation tools are wrapped as web services that receive and return TCF. Although it is not a strict requirement, many WebLicht web services use TCF as their input and/or output format, since its use increases the possibilities for combination with other tools in a workflow.

The WebLicht user interface provides users with a simple mechanism for creating and executing valid workflows. The results can be visualised in ways appropriate for individual annotation layers (e.g., tokens, lemmas, and part-of-speech tags in a table view; parse trees in a graphical view; named entities highlighted within the text).

The CLARIN Service Provider Federation is used by WebLicht to allow researchers to log in through their academic institutions; this makes WebLicht available to researchers from thousands of academic institutions.

Since its introduction in 2010, WebLicht has attracted a large user community, and is used in both research and teaching. As a usage indicator, WebLicht began counting tool invocations in 2014; WebLicht services have been invoked 2.5 million times in total, with an average of 30.000 invocations per month.

### 2.1.2 Language Resource Switchboard

The Switchboard enables interoperability by facilitating a seamless flow of data from all kinds of data hosts to many different types of tools that can process the data in one way or another (Zinn 2018b). The acknowledged strength of the Switchboard is based on very few assumptions. First, a resource can speak for itself, only a few of its intrinsic characteristics must be identified so that no externally-given metadata about the resource needs to be taken into account. Second, it is assumed that a tool can be characterised by the type of data it can process and the task it achieves with this processing. Third, data can reside anywhere *and* tools can be located anywhere as long as they are web-based, i.e., they respond to URL requests and provide a browser-based user interface.

The task of the Switchboard is to help users feed their data into tools that achieve tasks the users are interested in. By and large, the Switchboard acts as a simple broker. Once the Switchboard has received a data resource, the Switchboard profiles the resource's nature, and then identifies a list of *applicable* tools for such data, ordered by the tasks they can achieve. The users then select their tool of interest so that the Switchboard can forward the resource to the chosen web-based tool.

The Switchboard's design aims at maximising interoperability across many different services within and beyond CLARIN. Using the Switchboard GUI, data can be uploaded from a user's computer, referred to by a URL (e.g., by using the 'share file' function from Nextcloud or Dropbox), or generated on the fly by simply typing text. The Switchboard can also be invoked with a URL parameter that points to the location of data. Both the Virtual Language Observatory and the Virtual Collection Registry provide users with a button attached to data to invoke the Switchboard in this way. Also, the Switchboard has been integrated within a DataVerse-based data

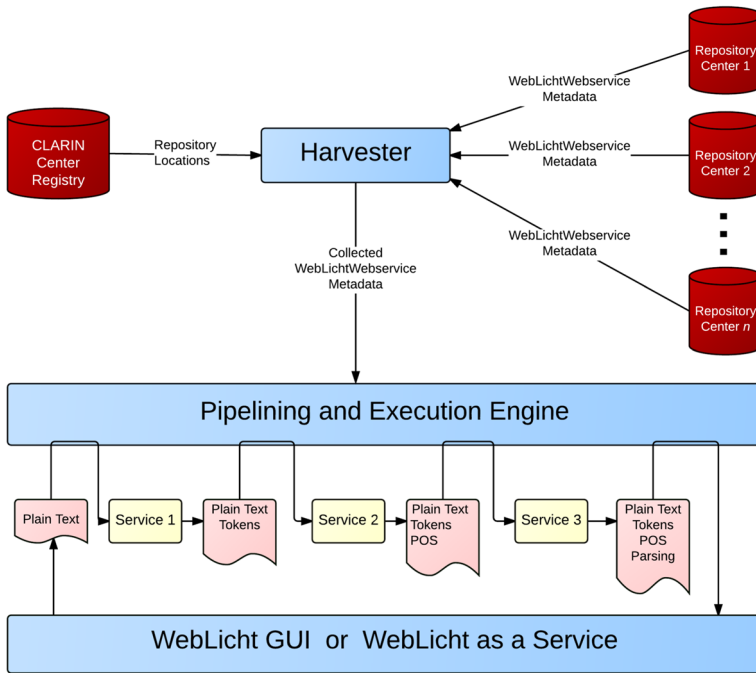


Fig. 1 WebLicht overview

repository.<sup>7</sup> Here, the Switchboard appears *inside* the GUI of the repository either as pop-up window or *iframe* embedding. Such integrations give users direct access to the Switchboard – without leaving the data site – and hence to its entire tool space, dramatically lowering the bar for tool discovery and use.

Since its inception, the Switchboard drives its tool selection by the mimetype of a resource (e.g., ‘text/plain’) and its language (e.g., ‘English’). Tools that can process, e.g., plain text files in English, are then shown in a task-based view. Five years on, it is still mime-type and language that drive tool selection. However, interoperability could be further improved by having the Switchboard (i) taking other resource characteristics into account, and (ii) changing the characteristics of a resource in order to increase the tool space that can process it. Consider the first case where the size of a resource may play a significant role in tool selection. A dependency parser for English may work perfectly fine for input data of a dozen sentences but fail miserably when given an entire novel. Here, tool applicability is a relative notion. The Tübingen team is currently developing a solution that is in line with the Switchboard’s design, hence off-loading the solution to the problem of large files to an external application. A new software on top of WebLicht, WebLicht-Batch<sup>8</sup> (Zinn and Campbell 2022), divides a given large file into smaller files, sends them to WebLicht for analysis, and constructs the result of the larger file by

<sup>7</sup> <https://dataverse.org>.

<sup>8</sup> <https://weblicht.sfs.uni-tuebingen.de>.

collecting the results of analysing the smaller files. Weblicht-Batch has been added to the Switchboard's tool space. By taking into account the size of a resource as an additional, third characteristic, the Switchboard will be able to identify WebLicht-Batch as 'most-suitable' applicable tool.

In its short life span, the Switchboard has built bridges across various infrastructures. Its simple design philosophy facilitated its integration with EUDAT's cloud space (see <https://b2drop.eudat.eu>; Zinn 2018a) and the D4Science platform (see <https://d4science.org>). Also, the Switchboard has been forked into an SSHOC-Switchboard that takes SSHOC-specific tools and tool preferences from the community into account. The Switchboard is also used in the evolving CLARIAH-DE infrastructure to bring together data and tools from its two predecessor infrastructures CLARIN and DARIAH, showcasing their successful merge by demonstrating interoperability across the two communities. In this respect, the Switchboard also plays a central role in the transatlantic LAPPS Grid project, alongside WebLicht.

### 2.1.3 LAPPS grid

The two-phase collaborative project *Transatlantic Collaboration between LAPPS and CLARIN* explores ways of integrating the Language Application Grid (LAPPS Grid) and CLARIN. The project partners include key members from LAPPS (Brandeis University and Vassar College) and CLARIN (University of Tübingen and Charles University in Prague).

LAPPS Grid is a web service infrastructure for language analysis, similar to WebLicht, but built on different underlying technologies.<sup>9</sup> In phase I of the project, interoperability between WebLicht and the LAPPS Grid workflow engine was successfully completed, allowing users of each workflow engine to use services from the other. The main challenges for this task included data conversion between the internal data formats used by each system, developing proxies to handle service requests using differing protocols, and conversion of tool metadata for use by the other system. In phase I of the project, a plan was also developed to allow mutual single-sign-on login to both systems by creating a trust network between CLARIN and the LAPPS Grid.

The main goals of phase II of the project are to implement the trust network according to the plan developed in phase I; to integrate the LAPPS Grid into the Switchboard; to integrate the UDPipe services developed at the CLARIN centre at Charles University into WebLicht and LAPPS Grid; and to implement a named entity tool that links named entities to unique identifiers from both authority records (for persons) and georeferences (for locations), and then stores the linked data in CMDI metadata records.

---

<sup>9</sup> <https://www.lappsgrid.org>.

## 2.2 Metadata schemas and data search

The CLARIN infrastructure encompasses a large space of language-related resources and tools and serves a diverse group of researchers from different communities. To connect users to data and tools relevant for their research, a rich set of metadata descriptors is needed. Each type of resources merits its own descriptors, because a lexical resource, for instance, is quite different in nature than a speech corpus, or a set of psycholinguistic experiments. For this reason, CLARIN decided to make use of a metadata framework rather than to resort to a fixed set of metadata schemas and descriptors.

The Component Metadata Infrastructure (CMDI)<sup>10</sup> is a metadata infrastructure that allows users to define metadata schemas, that is, *profiles* by grouping together predefined building blocks called *components*. Each metadata component is a structured collection of metadata fields or *data categories*. Each data category describes a particular aspect of a resource such as its type, genre or modality, or its speech rate and number of lexical entries, or for a software resource, its operating system or the programming language it supports for its application programming interface. It is important to note that a component can also contain other components so that hierarchically organised metadata schemas can be defined.

All components and profiles are stored in the *Component Registry*<sup>11</sup> and all elementary metadata fields are linked to data category registries giving them their semantic grounding.<sup>12</sup> CMDI-based metadata hence benefits from a semantic interoperability that helps both metadata designers and consumers. With CMDI being based upon XML technology, each profile can be converted into an XML Schema so that XML tools can be used to validate the completeness and correctness of metadata instances describing a particular resource of a given type.

At the time of writing, the component registry contains 182 profiles and 1260 components in the public and productive section. Users can define new profiles, or new components, whenever they feel that existing ones lack the expressiveness to properly describe their resources. Equally, new elementary descriptors can be defined in the CLARIN concept registry to semantically ground them.<sup>13</sup>

While the impressive number of profiles, components, and data categories in the registries indicates their broad user adoption, it also created the problem for users to identify the most appropriate ones for their needs (rather than defining their own ones). As a consequence, a CMDI task force has been investigating the use of metadata schemas and components in CLARIN practice, with the goal to identify

<sup>10</sup> <https://www.clarin.eu/content/component-metadata>, see also ISO 24622-1 and ISO 24622-2.

<sup>11</sup> <https://catalog.clarin.eu/ds/ComponentRegistry/>.

<sup>12</sup> In the beginning, most metadata descriptors were defined in the ISOcat registry, which was created in 2008 by the ISO/TC37 committee. Its management was later transferred to the Language Terminology/ Translation and Acquisition Consortium (LTAC) (see <http://datcatinfo.net/>), and finally replaced with the CLARIN Concept Registry (CCR), see <https://www.clarin.eu/ccr>.

<sup>13</sup> While the entries in the CLARIN Concept Registry are licensed via a CC-BY license (<https://creativecommons.org/licenses/by/4.0/>), entries in the CLARIN Component Registry are not explicitly licensed.



so-called *Core Components*.<sup>14</sup> The idea is to identify widely used components and recommend them for use for a good range of different application scenarios. CLARIN users are now asked to review those core components first before consulting (or adding to) the metadata registries.

Once users have selected (or defined) the CMDI profile that fits their data best, they can create instances of the profile to describe each resource. With CMDI being based on XML technology, the description of each resource can be validated, that is, it is tested whether an instance description supplies all mandatory metadata fields, or whether each value is of the correct datatype. In the spirit of CLARIN federation, most data providers store their data, together with their metadata descriptions, locally. To make such data findable in the entire CLARIN community, all local data is regularly harvested via the OAI-PMH<sup>15</sup> protocol and ingested into a central metadata catalogue, which all users can consult via the Virtual Language Observatory (VLO).<sup>16</sup> Besides CMDI-based metadata, the VLO is also able to ingest other metadata formats, for instance, those from DataCite<sup>17</sup> and the Open Language Archive Community<sup>18</sup> (OLAC).

The VLO is based on Apache Solr<sup>19</sup> and provides users with a faceted search over all harvested metadata. The facets that users can select to narrow down their search are derived from CMDI-based metadata fields. In fact, their common semantic grounding in the concept registry makes it possible to map a broad range of metadata fields to facets (Van Uytvanck et al., 2010), hence testifying to the semantic interoperability aspect of metadata. The VLO search interface is rich in options, including access to the most frequently occurring facet values, multiple facet value selection, combined full-text search on metadata with facet-based search, and the use of logical operations to combine elementary search queries to complex ones. Search results are displayed to cater for the most important user needs such as the availability of a resource, service or tool, and links to related metadata and their resources. An important feature of the VLO is its integration with other pillars of the CLARIN infrastructure. Users, for instance, can add metadata records to the collections they maintain in the Virtual Collection Registry, or they can process some textual resources directly with the Language Resource Switchboard.

The Virtual Language Observatory now gives CLARIN users access to over 1,000,000 language-related resources. The use of CMDI-based metadata, and the integration of the VLO with other tools (Switchboard, Collection Registry) makes it clear that the CLARIN community has found the technical means to implement the FAIR principles. Research data is easily findable (VLO), accessible (AAI, see section 2.3. for details), interoperable (CMDI-based metadata for the description of resources and tools), and re-usable (e.g., tool invocation within the VLO).

<sup>14</sup> <https://clarin-eric.github.io/cmd-i-core-components/>.

<sup>15</sup> <https://www.openarchives.org/OAI/openarchivesprotocol.html>.

<sup>16</sup> <https://vlo.clarin.eu>.

<sup>17</sup> <https://datacite.org>.

<sup>18</sup> <http://www.language-archives.org/>.

<sup>19</sup> <https://solr.apache.org/>.

### 2.3 Common infrastructure components for user authentication

The CLARIN platform gives users access to a vast space of language-related data and tools to process such data. All data and tools are geographically and organisationally distributed over CLARIN centres that often cross national boundaries. For a significant amount of data, however, intellectual property rights (e.g., newspaper corpora) or data protection issues (e.g., personal data from psycholinguistics studies) need to be taken into account, see section 3.2 for details. With a distributed infrastructure such as CLARIN, a well devised authentication and authorisation infrastructure (AAI) is the key for giving academic partners access to such data, but also to tools to process them.

CLARIN aims at making user authentication and authorisation as easy as possible. For this purpose, CLARIN has opted for a single sign-on approach where users can identify themselves with their university accounts. The login mechanism clearly identifies users (authentication), but also constitutes a first level of authorisation as it marks users as belonging to the academic community. With research data often licensed for academic use only, the single-sign on is often sufficient for most data and tool use. Access to data with more restrictive licences are granted on a case-by-case basis; here the AAI-based single sign-on serves at verifying a user's identify with technical means.

The interoperability of the authentication infrastructures of the thousands of institutions in the academic world has been a crucial element to ensure interoperability in research for quite some time, and it has led to the establishment of national identity federations such as DFN in Germany (<https://www.dfn.de/en/>), SURFnet in the Netherlands (<https://www.surf.nl/en>) or IDEM-GARR in Italy (<https://www.idem.garr.it/en/>). Each of these organisations is establishing a network of trust between its members (usually all or most universities and research organisations within the country) and together they have formed the European federation eduGAIN (<https://edugain.org/>), which in turn creates the technical setup for the establishment of trust between all identity providers (IdPs) and service providers (SPs) within the various national networks.

CLARIN has used this existing AAI infrastructure to achieve a high degree of interoperability within its network by establishing a so-called Service Provider Federation (SPF).<sup>20</sup> This is set up in such a way that CLARIN acts as a technical hub providing all the information about the various CLARIN centres (actually, each service provided by a centre to the CLARIN community) and making this information easily available to each national identity federation. CLARIN centres can register their services with CLARIN, using SAML-based metadata,<sup>21</sup> and CLARIN then distributes this metadata to all the identity federations of the participating countries. This makes it easier for the centres, as they do not have to ensure themselves that their metadata is propagated to all European identity federations, and on the other side CLARIN acts as a trusted party to the identity federation that guarantees the integrity of all registered services. With this setup,

<sup>20</sup> <https://www.clarin.eu/content/service-provider-federation>.

<sup>21</sup> [https://en.wikipedia.org/wiki/Security\\_Assertion\\_Markup\\_Language](https://en.wikipedia.org/wiki/Security_Assertion_Markup_Language).

the CLARIN AAI team can also help the centres with creating their metadata and making sure that they adhere to certain standards, e.g., the GÉANT code of conduct.<sup>22</sup>

Technically, CLARIN has set up a semi-automatic pipeline for adding and editing the SAML metadata definitions. CLARIN maintains a dedicated github repository<sup>23</sup> where all metadata definitions for all participating service providers are stored. Adding of new SPs and editing of existing ones is managed via pull requests to this repository. Each pull request triggers an automatic check of the metadata for consistency and completeness that will notify the submitter if anything is not correct. If all automatic checks are passed, the change is reviewed by the CLARIN AAI team and – if there are no issues – merged into the master branch and released to the public.

Additionally, CLARIN manages its own identity provider<sup>24</sup> that serves as a kind of *home for the homeless*. The IdP is meant to offer an easy way to use CLARIN resources and services for researchers that are not able to use their regular university account. There are various reasons for this, but the most common ones are twofold: either the researcher is temporarily not affiliated with any university or research institution, or the researcher's home organisation is, for whatever reason, not trusting or releasing the necessary user identification data to CLARIN SPs. The CLARIN IdP helps such researchers by granting them CLARIN accounts. Such accounts are only granted to academic users upon a well justified request reviewed by the CLARIN central office.<sup>25</sup>

## 2.4 Reproducibility of research results

Scientific knowledge is grounded on falsifiable predictions and thus its credibility and *raison d'être* rely on the possibility of repeating experiments and getting similar results as originally obtained and reported. Given the accelerated pace in research, the attention given to reproducibility in science has grown in recent years across all areas, with language science and technology being no exception in this trend (Branco 2013). Consequently, a number of tentative initiatives addressing reproducibility in this scientific area have been launched (Branco et al., 2018, 2017, 2016).

Given CLARIN's infrastructural mission in this field, one of its reaching-out initiatives concerns the support of novel and exploratory practices on matters which are not the immediate subject of research, but which are crucial to support scientific research and its vitality. One of these practices is the reproduction of research results, where CLARIN is teaming up with the European Language Resources Association (ELRA) and a first undertaking was the joint organisation of REPROLANG2020 (Branco et al., 2020). For REPROLANG, a new type of shared task was defined: it was partly similar to the usual competitive challenges, as all participants shared a

<sup>22</sup> <https://wiki.geant.org/display/eduGAIN/Data+Protection+Code+of+Conduct+Cookbook>.

<sup>23</sup> <https://github.com/clarin-eric/SPF-SPs-metadata>.

<sup>24</sup> <https://idm.clarin.eu/>.

<sup>25</sup> For details, see the corresponding FAQ on the CLARIN website at <https://www.clarin.eu/content/clarin-identity-provider>.

common goal; but it was also partly different to previous shared tasks, as its primary focus was on seeking support and confirmation of previous results, rather than on overcoming the state of the art with superior results.

The contribution of CLARIN to strengthening scientific reproducibility is motivated by its distinctive capacity to offer a service to the community both in terms of know-how and of equipment, which CLARIN was the main provider of during the above-mentioned pioneering undertaking. Additionally, many methodological innovations specifically targeting scientific reproducibility in language research and technology were advanced by contributions of CLARIN, as detailed in Branco et al. (2020).

As we are preparing further events along this path, the lessons learned indicate that, as far as reproducibility issues for the research on language are concerned, CLARIN is stepping forward and playing a much needed and fundamental service that is not within the mission or the reach of other types of organisations.

### 3 CLARIN as a knowledge hub

Research infrastructures based on distributed data and a federated service model have three main components:

1. a technical infrastructure, comprising both the technical facilities that provide users with access to data and tools, and the people that operate those facilities.
2. a set of commonly-agreed-upon organisational rules, measures and conventions that aim at ensuring a seamless interaction between infrastructure users, operators and components; this including the use of standards, access mechanism, usage licences, and quality assurance procedures.
3. a set of measures and facilities that aim at securing a continuous transfer of knowledge between all players involved in the construction, operation and use of the infrastructure; the players need knowledge and expertise to perform their tasks, and they generate new knowledge and expertise as they go along.

It is the third item which we refer to as the *CLARIN Knowledge Infrastructure*. It has been developed as the *glue*, holding together the various communities engaged with CLARIN, and as the structure that aims at securing a continuous transfer of knowledge and expertise between the diverse parties involved in the construction, operation and use of the infrastructure at large. The Knowledge Infrastructure has been built around a number of geographically distributed knowledge centres, offering and sharing expertise and best practices among the researchers in the network. The sharing of knowledge is also facilitated by a number of user involvement activities that are meant to facilitate the re-use of resources and the uptake of tools, standards and methodologies. Under the umbrella of the Knowledge Infrastructure, a rich number of online materials has been created that instruct users in using CLARIN to support their research effectively. Funding and other support instruments are also crucial to encourage exchange and adoption of shared practices. An effective

exchange is dependent on several aspects of organisational interoperability. In what follows, we will illustrate this point with some examples.

### 3.1 CLARIN Knowledge Centres

A fundamental part of the CLARIN Knowledge Infrastructure are the Knowledge Centres, or K-centres.<sup>26</sup> K-centres are institutions that have agreed to share their knowledge and expertise with others. A K-centre can be hosted at a single institution, or be distributed among various partner institutions. K-centres can be found in CLARIN member countries, but also exist elsewhere, and they all have a virtual presence. Each K-centre has a help desk and is committed to answering queries by users within two working days.

K-centres all have their own specific areas of expertise, focusing on individual languages (e.g., Czech, Danish, Polish, Portuguese), language families (e.g., South Slavic), or groups of languages (e.g., morphologically rich languages, the languages of Sweden). They can provide expertise on various modalities (e.g., written text, spoken language, sign language), or topics (e.g., language diversity, language learning, diachronic studies); they assist with language processing issues (e.g., speech analysis, building treebanks, machine translation), and various data types other than corpora (e.g., lexical data, wordnets and other lexical semantic networks, terminology banks). They can be contacted for information on how to use or process families of language data that will exist for most languages (e.g., newspapers, parliamentary records, oral history), and for generic methods and issues (e.g., data management, ethics, intellectual property rights, optical character recognition). K-centres are certified upon request by a special committee, for a period of three years; periodical reassessments of their activities are carried out.

With a growing number of K-centres being added to the infrastructure over the recent years, the question of making their expertise more easily accessible has arisen. A set of standardised metadata to describe each centre and its services has been developed, including areas of competence, audiences served, types of services, languages and modalities covered, linguistic and NLP topics, and data types. Users can use such descriptors to identify centres relevant for their research.

Links to other aspects of the CLARIN infrastructure have also been created, in particular connecting each centre to one or more Resource Families (curated collections of language resources) <https://www.clarin.eu/resourcefamilies>. K-centres were also inserted in the *Tour de CLARIN*<sup>27</sup> initiative, which enabled them to showcase their expertise and services, as well as the research that they helped to develop. At the time of writing, CLARIN is seeking to strengthen the collaboration among K-centres, using the aforementioned rich set of descriptors to encourage collaboration over specific areas (such as sign language), thus creating an interoperable system of knowledge sharing. In the future, a similar approach may

---

<sup>26</sup> <https://www.clarin.eu/content/knowledge-centres>.

<sup>27</sup> <https://www.clarin.eu/Tour-de-CLARIN>.

be extended to other aspects of the knowledge infrastructure, for instance, to the systematic creation and distribution of training materials.

### 3.2 Legal framework for interoperability

CLARIN offers its community access to a vast amount of language-related resources and tools that is distributed over 25 countries, or national consortia. It is not uncommon in the CLARIN community, say, to have a Dutch political scientist wanting to perform a study on post-war news coverage on the emerging influence spheres separating the West of Europe from the East of Europe. For this purpose, the scientist would like to get access to newspaper collections and TV news coverage from France, Germany, and Great Britain, but also from Poland, Czechoslovakia, and Finland between 1945 and 1955, all resources collected under different legal frameworks coming potentially under different licences. How could CLARIN help here?

The legal framework in the EU is intended to provide an interoperable space for various activities. This interoperability in legal framework can have two forms: harmonisation and unification. Harmonisation is introduced through EU Directives, which lay down certain minimum standards leaving the Member States the choice of exact measures to achieve them (in the process called *national implementation*). Intellectual Property Law, and especially Copyright Law, is highly harmonised at the EU level. Unification would be possible with Regulations, legal acts of the EU that apply directly in all EU Member States without the need for national implementation. As of now, no Regulations in the field of copyright (unlike in the field of data protection) exist.

Until recently, research was an area that was left to the discretion of the Member States. This still affects the sharing of research data and resources that can be achieved through a research infrastructure like CLARIN, as we need to find common legal ground that is applicable to research in all EU countries.

The intellectual property aspect of the legal space has been extensively discussed in Kelli et al. (2016, 2018a, 2019a) by members of the CLARIN Legal and Ethical Issues Committee. CLARIN recommends using Creative Commons licenses whenever possible (Oksanen and Lindén 2011). For all data sets, including those that cannot be made openly available, CLARIN offers a legal metadata classification system (Oksanen et al., 2010) to inform users of potential restrictions that they need to be aware of when accessing and working with a data set. For data sets that cannot be made openly and publicly available, CLARIN also offers standard license templates for depositing data to be shared through CLARIN Centres (Kelli et al., 2018b). The part of the intellectual property framework affecting the possibilities to share research data has been extensively scrutinised by CLARIN, and we are eagerly awaiting new opportunities provided by the EU Directive on Copyright in the Digital Single Market (Kelli et al., 2020b).

Despite the fact that all data cannot be made openly accessible, it is possible to use data to which one has lawful access for creating openly accessible language models. For a detailed discussion of this, see Kelli et al. (2020a).

During the last few years, the consequences of EU's General Data Protection Regulation (GDPR) has been widely recognised (Kelli et al., 2021). Some leeway was given to individual EU member countries to implement exceptions for research and this has led to differing practices for sharing personal data for academic research purposes (Kelli et al., 2019b; Lindén et al., 2020). In this context, it is particularly important to develop and promote best practices that can be re-used across CLARIN consortia, such as the one described below.

### 3.2.1 Use case: legal underpinnings for data collection in Finland

To illustrate how personal data can be collected and shared, we present the legal underpinnings of the *Donate Speech* campaign involving more than 25 000 citizens in Finland donating more than 220.000 speech samples comprising roughly 4000 h of colloquial speech.<sup>28</sup>

Since the definition of personal data in the GDPR is very broad, significant parts of the speech material is personal data for various reasons, as e.g., metadata about the speaker (such as his or her name) may be linked directly to an identifiable person. In addition, the recognisable voice of a speaker may also be linked to a person if there is some other information about the speaker available. The content of the speech may include personal information, e.g., the speaker reveals what he was doing with his friends last weekend.

According to the GDPR, *personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes*. Therefore, it was essential to define the purpose as specifically and clearly as possible, but without unnecessarily limiting the possibilities for future re-use of the data. Considering all this, the following definition was adopted: *Personal data is processed for the development and research of applications and services that understand and produce speech, as well as for language research and higher education related to these purposes*.

The processing of personal data is lawful only if one of the legal bases specified in Article 6 of the GDPR applies. In the described use case, choosing consent as a legal basis was considered impractical, because of the requirement for specificity and the possibility for the data subjects to withdraw their consents at any time. Instead, legitimate interest was chosen to be the most appropriate basis. However, if it becomes necessary to process special categories of personal data listed in Article 9 of the GDPR (e.g., political opinions, religious or philosophical beliefs or data concerning health or sex life) an explicit consent to the processing of such personal data is needed.

To inform the data subjects, i.e., the individuals who donate their speech to the campaign, two essential documents were drafted: (1) A *short information page*

<sup>28</sup> The practical background and the legal considerations of the particular collection campaign are more extensively elaborated upon in the CLARIN Book in the chapter on the campaign (Lindén et al., 2022).

including simple conditions of participation, which briefly describes the campaign and asks the person to accept the terms. (2) A comprehensive *data protection policy*, which aims to describe in a comprehensible and clear way how personal data are processed in the campaign, gives some basic information on data protection and describes the rights of the donor in order to secure the data subject's right to be informed.

The speech materials donated during the campaign are now stored in the Language Bank of Finland (Kielipankki), coordinated by the University of Helsinki as part of FIN-CLARIN. For academic researchers, the use of the data will be free of charge like the rest of the services of the Language Bank of Finland. For commercial use, a fee will be charged to cover handling costs.

#### 4 CLARIN community management

This section highlights how CLARIN promotes the adoption of common practices within and across its various communities. It provides a description of its user involvement and knowledge sharing activities at the central and national level and also cites the scientific disciplines involved.

CLARIN has a strong focus on action lines that aim at increasing the uptake of the CLARIN services available. For this purpose, it is taking into account the different types of users: providers (who produce resources and tools and need to store and disseminate their results), SSH scholars (who use CLARIN to do better, faster and more innovative research) and educators (who teach new generations of scholars with a thorough understanding of the added value of digital scholarship). Understanding the needs of these user categories in general, and the specific needs of the different research communities in particular, is seen as a precondition for meaningful demonstrations of the vital importance of language data for discovering new ways for machines to interact with humans, and for humans to interact with machines. More generally, insight into user needs is seen as the foundation for generating lasting impact, and for fostering innovation based on language resources.

User involvement is crucial for stimulating the uptake of CLARIN services and the transfer of knowledge. CLARIN has action lines aimed at these objectives both at the central and at the national level, with activities coordinated by the central office and initiatives from the various national consortia. Mobility grants, especially targeted at young researchers, as well as event grants to support the organisation of thematic events, also play an important role.

To cite but a few recent examples, 2021 saw a number of such initiatives: the DELAD workshop about sharing corpora of Speech with Communication Disorders, a CLARIN Café on Best Practices for Preserving Oral Archives, and one on Linguistic Linked Open Data, and also various training events. User involvement events are also important to collect information about possible gaps in terms of interoperability in the infrastructure, such as missing resources, incompatibility between tools and resources, or lack of alignment between similar resources in different languages. Once the needs have been identified, specific instruments can



be put in place to fill the gaps, in the form of funded projects, new initiatives or task forces, leading to the creation of new resources or the improvement of old ones.

For instance, the series of ParlaCLARIN workshops was crucial to establish the need for a set of comparable parliamentary corpora, which in turn led to the funding of the ParlaMint initiative.<sup>29</sup> In particular, such events allowed for the definition of guidelines for a common annotation scheme. The ongoing ParlaMint project is closely linked to the CLARIN Resource Families (RF) initiative, which addresses the need, expressed by DH researchers, for easy-to-use overviews of language resources organised by type. Parliamentary corpora were one of the first Resource Families, and the creation of this overview was instrumental to identify interoperability issues and missing corpora. More generally, a dedicated funding instrument is currently in place to support initiatives aimed at creating or improving resources, extending the ParlaMint model to other RFs. Finally the RF initiative, which involves the manual verification and curation of metadata, is crucial to identify possible interoperability issues at the LR description level. Missing or erroneous metadata are collected and centres are notified, so that the quality of metadata in the VLO can also be improved. More information on all of these initiatives can be found on the <https://www.clarin.eu/parlamint> portal.

Training initiatives can also be an important drive towards interoperability, in that they encourage the uptake of common standards and platforms within specific communities. It is evident that the uptake in the academic context is crucial for the dissemination of CLARIN resources and knowledge. An important sub-set of initiatives is aimed specifically at teachers and lecturers. The events organised thus far have proven to be of critical importance for identifying the needs in terms of training materials, and have led to the creation, now ongoing, of a CLARIN Learning Hub.<sup>30</sup>

#### 4.1 Multimodal resources in CLARIN-DK

One of the interoperability efforts in CLARIN has addressed the conversion of speech transcriptions from the CHAT format, used by the Conversation Analysis community, to the PRAAT textgrid format (Boersma and Weenink 2009), used by phoneticians and computational linguists. The more expressive PRAAT format makes it possible to automatically enrich transcriptions with information about pitch, intensity and formant contours, which in turn informs the analysis of the relation between speech and gestures. Conversion scripts were developed by the CLARIN K-centre at Carnegie Mellon University in the CLARIN-DK project. Researchers from the University of Copenhagen used the scripts to convert CHAT transcriptions of Danish spontaneous dyadic and triadic conversations collected and transcribed by researchers at the University of Southern Denmark to Praat<sup>31</sup> (MacWhinney and Wagner 2010). The Praat transcriptions were segmented

---

<sup>29</sup> <https://www.clarin.eu/parlamint>.

<sup>30</sup> <https://www.clarin.eu/content/learning-hub>.

<sup>31</sup> The audio- and video-recordings are available at <https://samtalebank.talkbank.org/>.

and time-stamped at the word level semi-automatically, the transcriptions were enriched with phonetic information, and the videos were annotated in ANVIL, a multimodal annotation tool popular in the software agent community (Kipp 2003). Gesture annotations, which comprise head movements, gaze, facial expressions, body posture and hand gestures, were coded according to the MUMIN annotation framework (Allwood et al., 2007).<sup>32</sup> The resulting multimodal annotations have been the base for studies addressing, e.g., the automatic prediction of speakers from their gestural behaviour (Navarretta 2014), the comparison of multimodal feedback in different communicative situations (Navarretta and Paggio 2012), transfer learning applied to the prediction of the functions of head movements in a multimodal corpus using training data from other corpora in the same language (Navarretta 2013) or in different languages (Navarretta and Lis 2014) containing annotations at different granularity levels.

In ongoing work under the Danish funded network Multimodal Child Language Acquisition,<sup>33</sup> researchers from the University of Copenhagen, the University of Hong Kong and the Chinese University of Hong Kong are combining and further developing the CHILDES-CHAT multilingual transcription conventions<sup>34</sup> and the MUMIN annotation scheme to account for multimodal phenomena in bilingual and trilingual child language acquisition (Matthews et al., 2021). PRAAT and the multimodal tool ELAN<sup>35</sup> are used in this work.

## 4.2 User-initiated services for literary studies

From the start, CLARIN has offered its users a good range of tools for basic text analysis. However, the availability of the tools does not automatically imply their use in practice. In the following, we describe tool use, development, adaption, and user adoption in the Polish CLARIN consortium (CLARIN-PL).<sup>36</sup>

During the initial phase of CLARIN-PL, while cooperating with selected *key users*, it has been noticed that many of their processing needs could already be addressed by existing tools. Putting the existing tools into action, however, required users organising them into proper tool chains and performing some standard format conversions (Piasecki 2014). The definition of tool chains, say, by using the LPMN chain definition language (Walkowiak 2018), blocked users from getting their data processed. With end-users taking the role of *intellectual sponsors*, or tool co-designers, and providing example data, research questions about the data, and possible data annotations to answer them, a potential tool space opened up. This inspired CLARIN-PL developers to populate this space by implementing a family of web-based research applications on top of CLARIN-PL services using tool chaining and format conversions. With this setting, it is not surprising that almost every

<sup>32</sup> The MUMIN annotation schemes for ANVIL and for ELAN are available at CLARIN-DK, see <http://hdl.handle.net/20.500.12115/43>.

<sup>33</sup> <https://cst.ku.dk/english/projects/multimodal-child-language-acquisition/>.

<sup>34</sup> <https://childes.talkbank.org/>.

<sup>35</sup> <https://archive.mpi.nl/tla/elan>.

<sup>36</sup> <http://clarin-pl.eu>.

single functionality covered in the new tool space can be traced back to a specific researcher who suggested it. Below, we discuss two tools that have been developed with this methodology: *LEM*<sup>37</sup> (Piasecki et al., 2018b) and *WebSty*<sup>38</sup> (Piasecki et al., 2018a).

*LEM* (Literary Exploration Machine) is a web-based application that provides users with a simple toolbox for carrying out basic statistical analyses of texts: a user inputs a set of text files, selects a processing task, and obtains its result as an Excel-based spreadsheet. The range of processing tasks is informed by incoming requests from users; it includes: lemmatisation (text to text), statistics of lemmatised proper names, entity graphs, word senses and their hypernyms, topic models, and verb characteristics. While *LEM* was originally targeted at literary scholars, end-user requests helped defining processing tasks that are also useful for psychologists, sociologists, media scholars, and scholars from other fields. The easy extension of *LEM* with new functionality benefited from the modular, distributed, and parallel architecture of the CLARIN-PL Language Technology Centre (LTC), see (Piasecki et al., 2017).

*Stylo*<sup>39</sup> (Eder et al., 2016) is a well-known Python package for stylometry analysis. However, its effective use requires programming skills so that some users fail to use it to its full extent. *WebSty* (Piasecki et al., 2018a) has been designed to be a service-less alternative. *WebSty* is implemented on top of the LTC architecture, thus it has immediate access to all the language tools as modules in a pipeline. This allows for flexibility in defining stylometric features and offers efficiency in processing even large volumes of data. The LTC architecture automatically scales up and down according to the incoming requests, all kept in a queue. Additional computing processes are started, run in parallel and closed according to the requests (for tools of higher computational cost, a larger number of processes is instantiated). *WebSty* works on the basis of clustering texts and text fragments. It is equipped with a rich set of visualisation means (developed in response to users' expectations) and tools for feature analysis, e.g., distinctive for different text clusters. *WebSty* was originally constructed only for Polish with the use of the best language tools for Polish, e.g., taggers or NERs, and only those that express good coverage and small error level. In response to users' interest in conducting research on materials in other languages than Polish, the application was first expanded with support for Hungarian (cooperation with HunCLARIN<sup>40</sup>) and next with six more languages (English, French, German, Spanish, Russian, and Hebrew). In the multilingual version<sup>41</sup> results of all tools are converted to Universal Tagset (Petrov et al., 2012) and to the UDPipe format (Straka and Straková 2017). A general processing service<sup>42</sup> based on spaCy<sup>43</sup> is provided as a default, but in case of all languages a

<sup>37</sup> <http://ws.clarin-pl.eu/lem.shtml>.

<sup>38</sup> <http://ws.clarin-pl.eu/websty.shtml>.

<sup>39</sup> <https://github.com/computationalstylistics/stylo>.

<sup>40</sup> <https://clarin.hu/en>.

<sup>41</sup> <http://ws.clarin-pl.eu/webstym1.shtml?en>.

<sup>42</sup> <http://ws.clarin-pl.eu/spacy.shtml>.

<sup>43</sup> <https://spacy.io>.

set of tools performing the best in stylometric analysis of the given language has been identified with the help of interested users or the team of the given CLARIN partner. For all supported languages, a test set of literary works was also collected to evaluate a given language version of the application. WebSty was next used as a basis for development of research tools for topic modelling and clustering-based semantic analysis of text collections.

### 4.3 Digital Humanities in Germany

There is a very active Digital Humanities (DH) community in Germany that spans over a wide range of humanities disciplines. In 2012, DH scholars from the German-speaking countries formed their own professional association DHd (short for: *Digital Humanities im deutschsprachigen Raum*), which currently has more than 400 members and which organizes annual scientific meetings with large numbers of participants. At present, their research infrastructure needs are served by the NFDI4Culture and Text+ consortia, which are funded under the National Research Data Infrastructure (NFDI) program of the DFG, as well as by the CLARIAH-DE, CLARIN-D, and DARIAH-DE research infrastructure consortia. Additional NFDI consortia, NFDI4Objects and NFDI4Memory, have been approved in fall 2022 and are expected to start in early 2023. From the very start of the CLARIN-D project in 2006, disciplinary working groups in the areas of Linguistics, Literary Studies, Cognitive Psychology, History, Political and Social Sciences, and various Philologies were formed with the aim of helping to articulate the research needs of the scientific communities involved, to organise DH training events, to prioritise the curation of digital data and of software services, and to integrate such resources into the distributed network of CLARIN-D Centres. Such curation activities led, *inter alia*, to the compilation of the GERPAPARL corpus of parliamentary protocols (Blätte and Blessing 2018) and to the annotation tool WebAnno (Yimam et al., 2013; de Castilho et al., 2014) for the manual annotation and postediting of corpus data. Both resources are widely used in Germany and many other countries, and have been made interoperable with other annotation tools such as WebLicht and with other data sets such as the corpus collection developed by the PARLAMint initiative.

In the area of training, members of the CLARIN-D initiative have regularly offered training events for young researchers to familiarise them with the use of CLARIN-D corpus resources and associated software tools and services. On a regular basis, the CLARIN-D consortium has also contributed training courses to the scientific program of the European Summer University in Digital Humanities (ESU), organised by Elisabeth Burr and her team and held annually for many years at the University of Leipzig.

### 4.4 Digital Humanities in Poland

In 2005, when the CLARIN project was conceived, the Polish language was considered a low-resourced language; it had only a few corpora of limited extent,

a single tagger, a small wordnet, but also a good-coverage morphological analyser. As a result, there was little SSH scholars could use to drive their research with the computer-supported analysis on Polish texts. Thus, the main CLARIN objectives for the Polish national consortium were two-fold, the development of robust language technology (LT) for Polish, and the building of a community of SSH researchers interested in the development of the digital paradigm. From the very beginning, CLARIN-PL members subscribed to the open science paradigm (later known as FAIR principles) and a consortium of six partners was established in 2006.<sup>44</sup> The common denominator of the six partners was the shared understanding that all language resources and tools (LRTs) shall be free. Benefiting from the funding of several projects at national and European level, CLARIN-PL had managed to significantly improve the state of Polish LT before the 2013 launch of the CLARIN-PL construction phase – Poland is a CLARIN member since 2012.

In 2013, a bidirectional model for the development of Polish LT research infrastructure was proposed (Piasecki 2014), on the one hand, the provision of access to NLP-based research tools, and on the other hand, the further development of the tool space to fill-in the gaps for both resources coverage and tool functionality. A B-type CLARIN centre offers users a data repository and serves as an access point to central services and web-based tools. With regard to the tool space, the BLARK proposal (Krauwier 2003) was chosen as a reference point for the development of a suite of fundamental Polish LRTs, with a clear emphasis on tools that SSH researchers need, see Piasecki (2014).

To better identify users' needs, CLARIN-PL invited researchers familiar with, or interested in, digital methods to become *key users*. CLARIN-PL asked those users for tool functionality they wish to see realised, and offered person-months to work together on tool prototypes while they were sharing their expertise, time and data with developers. Starting from a few seed-users, the network of key users has been continuously growing since then, very often on the basis of individual recommendations. Once the first tool prototypes became available, CLARIN-PL initiated hands-on training workshops at users' home institutions around Poland. While the first events had a rather large audience ( $\approx 80$  participants), later events aimed at more manageable audiences, with a better focus on specific domains, groups of researchers, or processing tasks. The most fruitful workshops were those where participants could play with the new tools using their own data, fuelled by their specific research interests. In addition to the training aspect, it is worth to note that, very often, such events informed the further development of the tools themselves. Listening to end-users, the BLARK-guided development of LRTs was revised, a further development of deep parsers was stopped, and more development resources were shifted towards language modelling and distributional semantics (e.g., different forms of embeddings).

On the grander scale, CLARIN-PL supported efforts in establishing the Polish national node of DARIAH. Most of the DARIAH-PL members are CLARIN-PL users, and in the DARIAH-PL investment project (started in 2021), CLARIN-PL

<sup>44</sup> <https://clarin-pl.eu>.

plays the role of LT supplier with the focus on building technological interfaces linking it with this emerging new infrastructure.

#### 4.5 Digital Humanities in Latvia

Many fundamental LRTs are well known and widely used by almost every citizen of Latvia. Consider, for instance, <https://tezaurs.lv>, which has become an indispensable and widely cited Latvian dictionary (Spektors et al., 2016, 2019). However, many SSH researchers in Latvia still have insufficient knowledge on how NLP techniques can help to solve specific research questions. CLARIN-LV addresses this issue by organising dedicated seminars and tutorials<sup>45</sup> for them through the CLARIN SAF-MORIL K-Centre.<sup>46</sup>

The goal of supporting and educating Digital Humanities researchers has also been pursued by the Digital Humanities Initiative in Latvia.<sup>47</sup> Since 2020, this initiative is supported through several collaborative projects, for example, Language Technology Initiative project of Recovery and Resilience Facility and several projects of the state-funded research programme “Digital Resources for Humanities: Integration and Development”, with the CLARIN-LV consortium as a project partner. These projects support a wider application and further development of the existing digital resources and tools specifically for the Humanities to advance Digital Humanities research and education in Latvia. They provide a balanced and diverse programme, including educational initiatives, integration of similar humanities resources to prevent their fragmentation, as well as carrying out an analysis of the use and the users of Digital Humanities tools and resources to understand their effectiveness and to suggest ways to improve their usability for different target groups.

The cooperation and synergy with CLARIN-LV was established not only through the creation and adaptation of LRTs, but also by the prevention of tool fragmentation or unneeded duplicatory work, as well as by ensuring wider visibility and improved public access. While these projects cover many different activities, in this subsection we highlight two indispensable LRTs that have been adapted and customised for use in Digital Humanities:

- *korpuss.lv* – a central language corpus platform in Latvia. Our aim is to make the corpus platform available to any research group or individual researcher/student by supporting not only querying the available Latvian text corpora, but also allowing to process and deploy new text corpora on the shared platform (Saulite et al., 2022).
- NLP-PIPE<sup>48</sup> – a scalable CLARIN integrated NLP pipeline as a service for Latvian, supporting the annotation of Latvian texts both grammatically

<sup>45</sup> Seminar materials available at: <https://www.clarin.lv/lv/clarin-latvija-seminari>.

<sup>46</sup> <https://www.kielipankki.fi/safmoril/>.

<sup>47</sup> <http://digitalhumanities.lv>.

<sup>48</sup> <http://nlp.ailab.lv>

and semantically (Znotins and Cirule 2018). Its capability for named entity recognition is being integrated into the Latvian literature platform,<sup>49</sup> allowing the automatic identification of person names, place names, and events mentioned in texts; different Latvian NLP tools are being applied on the text analysis platform of the Latvian National Digital Library.<sup>50</sup>

## 4.6 Beyond academia

CLARIN has industrial-strength solutions for the archiving, processing, and automatic enrichment of language resources that may be also applied to complex multimodal resources that involve or are linked to language resources. Thus, CLARIN's relevance or impact goes beyond traditional research in the Social Sciences and Humanities and encompasses the entire *GLAM* sector (galleries, libraries, archives, and museums). CLARIN has a long-standing cooperation with Europeana,<sup>51</sup> a web portal created by the European Union that gives users access to digitised cultural heritage collections of more than 3,000 institutions across Europe. Some of Europeana's metadata, for instance, is ingested into the CLARIN VLO. At the national level, CLARIN-PL investigated the use of its named entity recognition service for the augmentation of documents' metadata in the Polish National library.

CLARIN-PL expertise and NLP services have also been put to use in a number of public service institutions:

- Services for sentiment analysis and the semantic analysis of texts based on topic modelling and clustering are used by the Centre For Strategic Analysis of The Chancellery of the Prime Minister of Poland.
- Services and resources (lexical semantic resources and word embeddings) have been developed in cooperation with the Educational Research Institute in Warsaw for the processing of qualification descriptions.
- A prototype for the automated recognition of abusive clauses in consumer agreements has been built for the Polish Office of Competition and Consumer Protection; this including the provision of training and test data.

CLARIN's expertise also attracts interest from business partners. Since 2013, tools from CLARIN-PL, for instance, have been used by more than 100 industrial users, raising the prospect for CLARIN-PL and other CLARIN consortia to systematically seek and profit from industry partnerships. In Poland, the CLARIN-PL-Biz project<sup>52</sup> addresses the issue with a two-legged approach: it will spend 60% of its resources with the aim at providing open research for all, and for the remaining 40%, it aims at securing industry contracts to deliver custom-made infrastructural LT services

---

<sup>49</sup> <https://literatura.lv/>.

<sup>50</sup> <https://lndb.lv/>.

<sup>51</sup> <https://www.europeana.eu>.

<sup>52</sup> <https://clarin.biz>.

to industry. The CLARIN-PL-Biz project has started with an initial investment of around 30 million Euros. Its main objectives are the further development of language resources and tools for Polish (and their linkage with other Slavic languages but also English) so that they can be used in a robust manner. Furthermore, it is planned to build a set of infrastructural LT services for selected application domains such as the training of speech corpora, dialogue analysis, semantic indexing and processing, contextual question-answering as well as personalised sentiment and emotion analysis. Nearly a third of the initial investment sum will fund the construction of a high-performance computing cluster and a repository centre for data storage. So far, CLARIN-PL-Biz obtained almost 4.4 million Euros worth of in-kind contributions (e.g., in terms of software licences and person-months) from more than 30 business partners, who also contribute in terms of LRT requirements analysis. It is hoped that the partner's involvement transforms into commercial products, whose technological development feeds back into the open science arm of the project.

## 5 Related work

The CLARIN research infrastructure offers its community an interoperable LT platform with access to a wide range of geographically distributed NLP tools and data, and a knowledge hub consisting of a growing number of K-centres providing users with knowledge in many areas of expertise. Community building is a central, ongoing task in CLARIN, and is performed in strong cooperation with CLARIN's national consortia. With its focus on language-related resources and tools, CLARIN complements research infrastructures in Europe that have a different or wider scope.

Other infrastructure projects focused on language data include the European Language Grid (ELG) and European Language Resources Coordination (ELRC). ELG<sup>53</sup> provides a scalable cloud platform for the entire European LT community, including research and industry, enabling providers, developers, integrators and consumers to share services, tools, products, datasets and other language resources (Rehm 2023). The platform also enables exchange of standardised metadata records. ELRC<sup>54</sup> is an initiative aimed at managing the relevant language resources in all official languages of the EU Member States, in order to help improve the quality, coverage and performance of automated translation solutions in the context of current and future digital services (Lösch et al., 2018). ELRC has also set up a network of experts, called National Anchor Points. Language resources collected or developed (mainly through focused web crawling) within the ELRC framework are made available via the ELRC-SHARE repository<sup>55</sup>(Piperidis et al., 2018). Some of the institutions and researchers participating in the two projects are also members of CLARIN, and they also provide some of the resources distributed in CLARIN via the two projects' platforms demonstrating the reusability of these resources

<sup>53</sup> <https://live.european-language-grid.eu>.

<sup>54</sup> <https://www.lr-coordination.eu/>.

<sup>55</sup> <https://www.elrc-share.eu>.



and following the strategy in Open Science that resources and services should be searchable in different contexts and reusable by different communities.

DARIAH is the research infrastructure closest to CLARIN.<sup>56</sup> Also, pan-European by design, it aims at supporting digitally-enabled research and teaching across the Arts and Humanities. Compared to CLARIN, which has a strong focus on Linguistics and NLP processing tools, DARIAH's target audience is hence broader in scope. DARIAH's Social Sciences & Humanities Open Marketplace,<sup>57</sup> for instance, lists around 1600 tools and services, including around 160 tools that can be used for annotating data (including, for instance, the WebLicht workflow engine and the Switchboard), but also many other tools in other areas. In part, the overlap can be explained by the CLARIAH-DE<sup>58</sup> and CLARIAH-NL<sup>59</sup> projects, which resulted from a merger of the respective national infrastructures in Germany and in The Netherlands.

The CESSDA consortium is another European Research Infrastructure Consortium.<sup>60</sup> for the Social Sciences. CESSDA has a stronger focus on data and metadata, aiming at providing services to support researchers (data producers) to describe and store their data. Its main service for finding relevant data sets is the CESSDA Data Catalogue, which mirrors the functionality of the CLARIN VLO. For the description of data, it provides a Vocabulary Service of controlled vocabularies (similar to the CLARIN concept registry) and the ELSST Thesaurus.

In the Life Sciences, the scientific workflow system Galaxy is very prominent and widely used.<sup>61</sup> While in principle largely domain agnostic, the Galaxy tool space is dominated by tools from the areas of metagenomics, proteomics, genome assembly and annotation, and from other areas in bioinformatics. With bioinformatics being data driven by nature, Galaxy also offers tools and workflows for statistical analyses and machine learning, and also provides access to interactive environments such as R<sup>62</sup> and Jupyter notebooks.<sup>63</sup> The Galaxy project in a global undertaking with a strong presence in Europe. It offers its growing community a good collection of training materials and organises regular training workshops. The European community is engaged in several projects at European, national, and regional level. In the US, the Language Application Grid provides its own instantiation of a Galaxy server, offering its users scientific workflows for NLP analyses.

---

<sup>56</sup> <https://www.dariah.eu>.

<sup>57</sup> <https://marketplace.sshopencloud.eu>.

<sup>58</sup> <https://www.clariah.de>.

<sup>59</sup> <https://www.clariah.nl/>.

<sup>60</sup> <https://www.cessda.eu/>.

<sup>61</sup> <https://galaxyproject.org>.

<sup>62</sup> <https://www.r-project.org>.

<sup>63</sup> <https://jupyter.org>.

## 6 Concluding remarks

The CLARIN consortium subscribes to the Open Science agenda where access to resources is provided adhering to the FAIR principles for data management. As underlined by several prominent bodies such as the European Commission (Collins et al., 2018), the added value of the Open Science agenda and the underlying values are to be sought in the way they stimulate research communities to work towards the realisation of the objectives, rather than considering features such as interoperability as an absolute condition for acceptance by the ecosystem. Interoperability can only be pursued effectively if it is not just targeted at technical levels; technology must be embedded in a culture that is characterised by attention to all social, political, and organisational factors that impact system-to-system performance.

Interoperability is at the heart of the CLARIN infrastructure. At the technical level, interoperability allows CLARIN to effectively operate a distributed network of data and tool providers. A common metadata framework, with semantic interoperability at its heart, is crucial for the Virtual Language Observatory to give users access to a large and diverse set of resources. CLARIN's work on common vocabularies and standards, its consideration of legal issues and licences makes sure that resources and their data can be used across national boundaries. The technical infrastructure, however, must be complemented with a knowledge infrastructure. The education of CLARIN users in many different aspects is a precondition for an effective use of the technical infrastructure. Users (and different user groups) need to be aware of (i) the tools that exist in the community (e.g., the Switchboard helps users to explore the CLARIN tool space given their actual data); (ii) the huge amounts of research data already available and ready for reuse (e.g., the VLO to explore all kinds of language-related research data); and (iii) the licences CLARIN promotes to publish their research data in CLARIN repositories so that they can be reused by others. All this includes basic knowledge about CMDI so that research data and tools can be described as concise and precise as possible. Training activities as carried out by the 25+ K-centres make sure that interoperability in the technical sense of the term carries over to interoperability on the mindset of the CLARIN community. An attractive technical infrastructure in place, with users being knowledgeable about the effective use of resources and tools, together with all kinds of events and measures to support community building, will automatically yield an increasing user community that benefits from the infrastructures, contributes to them, and hence ensures CLARIN's sustainability for the many years coming.

**Acknowledgements** The construction of the CLARIN infrastructure officially started on 29 February 2012, when the CLARIN ERIC (European Research Infrastructure Consortium) was created, with nine founding members. CLARIN ERIC's main task is to build, operate, coordinate and maintain the CLARIN infrastructure; it neither conducts nor funds research activities. CLARIN ERIC was set up with financial support from the European Commission through the CLARIN Preparatory Phase Project (2008–2011) – Funded by FP7-INFRASTRUCTURES, Grant agreement ID: 212230. It is now entirely funded by the participating countries. The results reported here were partly supported in Portugal by Lisboa2020, Alentejo2020 and FCT-Fundação para a Ciência e Tecnologia under the grant PINFRA/22117/2016 for PORTULAN CLARIN. The results reported here have been achieved under the project LUSyD, supported by the Czech Science Foundation, project No. GX20-16819X, and by the LINDAT/CLARIAH-CZ Research Infrastructure project, supported by the Ministry of Education, Youth and Sports of the

Czech Republic, under projects No. LM2018101 and LM2023062. This study has been supported by the EU Recovery and Resilience Facility project Language technology Initiative (No 2.3.1.1.i.0/1/22/I/CFLA/002). The reported work has been partially supported by CLaDA-BG, the Bulgarian National Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH, Grant Number DO1-301/17.12.21.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Allwood, J., Cerrato, L., Jokinen, K., Navaretta, C., & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3–4), 273–287. <https://doi.org/10.1007/s10579-007-9061-5>
- Blätte, A., & Blessing, A. (2018). The GermaParl Corpus of Parliamentary Protocols. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18). European Language Resources Association, Miyazaki, Japan, <https://aclanthology.org/L18-1130.pdf>
- Boersma, P., & Weenink, D. (2009). Praat: doing phonetics by computer (version 5.1.05). <http://www.praat.org>
- Branco, A. (2013). Reliability and meta-reliability of language resources: Ready to initiate the integrity debate? In: Kuebler S, Osenova P, Volk M (eds) Proceedings of TLT2013—12th Workshop on Tree-banks and Linguistic Theories, Bulgarian Academy of Science, pp 27–36, <http://www.di.fc.ul.pt/~ahb/pubs/2013bBranco.pdf>
- Branco, A., Calzolari, N., & Choukri, K. (2016). 4REAL Proceedings of the Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language. European Language Resources Association, [http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-4REAL\\_Proceedings.pdf](http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-4REAL_Proceedings.pdf), collocated with The 10th International Conference on Language Resources and Evaluation (LREC'16)
- Branco, A., Calzolari, N., & Choukri, K. (2018) 4REAL2018 Proceedings of the 2nd Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language. European Language Resources Association, [http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-4REAL\\_Proceedings.pdf](http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-4REAL_Proceedings.pdf), collocated with The 11th International Conference on Language Resources and Evaluation (LREC'18)
- Branco, A., Calzolari, N., Vossen, P., Van Noord, G., van Uytvanck, D., Silva, J., Gomes, L., Moreira, A., & Elbers, W. (2020) A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with reprolang2020. In: Proceedings of The 12th Language Resources and Evaluation Conference (LREC'20). European Language Resources Association, Marseille, France, pp 5539–5545, <https://www.aclweb.org/anthology/2020.lrec-1.680>
- Branco, A., Cohen, K. B., Vossen, P., Ide, N., & Calzolari, N. (2017). Replicability and reproducibility of research results for human language technology: Introducing an LRE special section. *Language Resources and Evaluation*, 51(1), 1–5. <https://doi.org/10.1007/s10579-017-9380-0>
- Collins, S., Genova, F., Harrower, N., Hodson, S., Jones, S., Laaksonen, L., Mietchen, D., Petrauskaitė, R., & Wittenburg, P. (2018). *Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data.*

- de Castilho, R. E., Biemann, C., Gurevych, I., & Yimam, S. M. (2014). Webanno: A flexible, web-based annotation tool for CLARIN. In: Proceedings of the CLARIN Annual Conference 2014. CLARIN ERIC, <http://tubiblio.ulb.tu-darmstadt.de/98002/>
- Dima, E., Hinrichs, E., Hinrichs, M., Kislev, A., Trippel, T., & Zastrow, T. (2012). Integration of WebLicht into the CLARIN Infrastructure. In: Proceedings of the Joint CLARIN-D/DARIAH Workshop at Digital Humanities Conference 2012: Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts, Hamburg, Germany, pp 17–23, <https://ids-pub.bsz-bw.de/frontdoor/index/index/year/2022/docId/10869>
- Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: A package for computational text analysis. *The R Journal*, 8(1), 107–121. <https://doi.org/10.32614/RJ-2016-007>
- Fišer, D., Lenardič, J., Erjavec, T. (2018). CLARIN's key resource families. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC'18). European Language Resources Association, Miyazaki, Japan, <https://aclanthology.org/L18-1210>
- Fišer, D., & Witt, A. (Eds.). (2022). *CLARIN: The Infrastructure for Language Resources*. Berlin, Boston: De Gruyter. <https://doi.org/10.1515/9783110767377>
- Heid, U., Schmid, H., Eckart, K., & Hinrichs, E. (2010) A corpus representation format for linguistic web services: The D-SPIN text corpus format and its relationship with ISO standards. In: Proceedings of the seventh international conference on language resources and evaluation (LREC'10). European Language Resources Association, Valletta, Malta, [http://www.lrec-conf.org/proceedings/lrec2010/pdf/503\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2010/pdf/503_Paper.pdf)
- Hinrichs, M., Zastrow, T., Hinrichs, E. (2010). WebLicht: Web-based LRT services in a distributed eScience infrastructure. In: Proceedings of the seventh international conference on language resources and evaluation (LREC'10). European Language Resources Association, Valletta, Malta, [http://www.lrec-conf.org/proceedings/lrec2010/pdf/270\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2010/pdf/270_Paper.pdf)
- ISO 24622-1. (2015). Language resource management—Component Metadata Infrastructure (CMDI)—Part 1: The component metadata model. International Standard, International Organization for Standardization (ISO), Geneva, <https://www.iso.org/standard/37336.html>
- ISO 24622-2. (2019). Language resource management – Component Metadata Infrastructure (CMDI)—Part 2: The component metadata specification language. International Standard, International Organization for Standardization (ISO), Geneva, <https://www.iso.org/standard/64579.html>
- Kelli, A., Lindén, K., Tavast, A., Vider, K., Birštonas, R., Labropoulou, P., Kull, I., Tavits, G., & Värvi, A. (2019a). The extent of legal control over language data: the case of language technologies. In: Proceedings of CLARIN annual conference 2019, [https://www.clarin.eu/sites/default/files/clarin2019\\_p4\\_20\\_kelli\\_tavast\\_linden\\_vider\\_birstonas\\_labropoulou\\_kull\\_tavits\\_varv.pdf](https://www.clarin.eu/sites/default/files/clarin2019_p4_20_kelli_tavast_linden_vider_birstonas_labropoulou_kull_tavits_varv.pdf)
- Kelli, A., Lindén, K., Vider, K., Kamocki, P., Birštonas, R., Calamai, S., Labropoulou, P., Gavriilidou, M., Stranák, P. (2019b). Processing personal data without the consent of the data subject for the development and use of language resources. In: I. Skadina and M. Eskevich M (eds), Selected papers from the CLARIN Annual Conference 2018, Pisa, 8–10 October 2018. Linköping University Electronic Press, Sweden, no. 159 in Linköping Electronic Conference Proceedings, pp 72–82, <https://www.clarin.eu/event/2018/clarin-annual-conference-2018-pisa-italy>
- Kelli, A., Lindén, K., Vider, K., Kamocki, P., Tavast, A., Birštonas, R., Tavits, G., Keskküla, M., Labropoulou, P., Kull, I., Värvi, A., Erikson, M., Vutt, A., & Calamai, S. (2021). Sharing is caring: a legal perspective on sharing language data containing personal data and the division of liability between researchers and research organisations. In: Selected Papers from the CLARIN Annual Conference 2020. Virtual Event, 2020, 5–7 October, Linköping University Electronic Press, pp 129–147, <https://doi.org/10.3384/ecp18015>
- Kelli, A., Mets, T., Vider, K., Värvi, A., Jonsson, L., Lindén, K., & Birštonas, R. (2018a). Challenges of transformation of research data into open data: The perspective of social sciences and humanities 1. *International Journal of Technology Management and Sustainable Development*, 17(3), 227–251. [https://doi.org/10.1386/tmsd.17.3.227\\_1](https://doi.org/10.1386/tmsd.17.3.227_1)
- Kelli, A., Lindén, K., Vider, K., Labropoulou, P., & Ketzan, E. (2018b). Implementation of an open science policy in the context of management of CLARIN language resources: A need for changes?, Linköping Electronic Conference Proceedings, vol 147, Linköping University Electronic Press, pp 102–111. [https://www.clarin.eu/sites/default/files/Kelli-et-al-CLARIN2017\\_paper\\_27.pdf](https://www.clarin.eu/sites/default/files/Kelli-et-al-CLARIN2017_paper_27.pdf)
- Kelli, A., Tavast, A., Lindén, K., Birštonas, R., Labropoulou, P., Vider, K., Kull, I., Tavits, G., Värvi, A., & Mantrov, V. (2020a). Impact of legal status of data on development of data-intensive products: Example of language technologies. *Legal Science: Functions, Significance and Future in Legal Systems II*. <https://doi.org/10.22364/isfcl.7.2.31>


- Kelli, A., Tavast, A., Lindén, K., Vider, K., Birštonas, R., Labropoulou, P., Kull, I., Tavits, G., Värvi, A., Straňák, P. and Hajič, J. (2020b) The impact of copyright and personal data laws on the creation and use of models for language technologies. In: Selected Papers from the CLARIN Annual Conference 2019, Linköping University Electronic Press, <https://ep.liu.se/ecp/172/008/ecp20172008.pdf>
- Kelli, A., Vider, K., Lindén, K. (2016). The regulatory and contractual framework as an integral part of the clarin infrastructure. In: Proceedings of the CLARIN Annual Conference 2016, <https://ep.liu.se/ecp/123/002/ecp15123002.pdf>
- Kipp, M. (2003) Gesture generation by imitation: From human behavior to computer character animation. PhD thesis, Saarland University, <https://doi.org/10.22028/D291-25852>
- Krauer, S. (2003) The basic language resource kit (blark) as the first milestone for the language resources roadmap. In: Proceedings of SPECOM, pp 8–15, <http://www.elsnet.org/dox/krauer-specom2003.pdf>
- Lindén, K., Jauhianinen, T., Lennes, M., Kurimo, M., Rossi, A., Kurki, T., Pitkänen, O. (2022). Donate speech—collecting and sharing a large-scale speech database for social sciences, humanities and artificial intelligence research and innovation. In D. Fišer and A. Witt (Eds.), *The CLARIN book*. Berlin: de Gruyter. <https://doi.org/10.1515/9783110767377-019>
- Lindén, K., Kelli, A., & Nousias, A. (2020). A CLARIN contractual framework for sharing personal data for scientific research. In: Selected Papers from the CLARIN Annual Conference 2019, Linköping University Electronic Press, [https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/10081/file/Kelli\\_Linden\\_Vider\\_Kamocki\\_et\\_al\\_CLARIN\\_contractual\\_framework\\_2020.pdf](https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/10081/file/Kelli_Linden_Vider_Kamocki_et_al_CLARIN_contractual_framework_2020.pdf)
- Lösch, A., Mapelli, V., Piperidis, S., Vasiljevs, A., Smal, L., Declerck, T., Schnur, E., Choukri, K., & van Genabith, J. (2018). European language resource coordination: Collecting language resources for public sector information management. In: N. Calzolari, K. Choukri, C. Cieri, et al. (eds) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan, <https://aclanthology.org/L18-1213.pdf>
- MacWhinney, B., & Wagner, J. (2010). Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gesprächsforschung Online-Zeitschrift zur verbalen Interaktion*, 11, 154–173.
- Matthews, S., Navarretta, C., Paggio, P., Ping Ping Tse, A. & Yip, V. (2021). Towards the construction of multimodal bilingual child language acquisition corpora. In: Second International Workshop on Multimodal Language Acquisition., University of Copenhagen
- Navarretta, C. (2013) Transfer learning in multimodal corpora. In: IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom), pp 195–200, <https://doi.org/10.1109/CogInfoCom.2013.6719240>
- Navarretta, C. (2014). The automatic identification of the producers of co-occurring behaviours. *Cognitive Computation*, 6(4), 689–698. <https://doi.org/10.1007/s12559-014-9269-9>
- Navarretta, C., & Lis, M. (2014) Transfer learning of feedback head expressions in Danish and Polish comparable multimodal corpora. In: Proceedings of the ninth international conference on language resources and evaluation (LREC'14). European Language Resources Association, Reykjavik, Iceland, pp 3597–3603, [http://www.lrec-conf.org/proceedings/lrec2014/pdf/525\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/525_Paper.pdf)
- Navarretta, C., & Paggio, P. (2012) Verbal and non-verbal feedback in different types of interactions. In: Proceedings of the eighth international conference on language resources and evaluation (LREC'12). European Language Resources Association, pp 2338–2342
- Oksanen, V., & Lindén, K. (2011) Open content licenses—How to choose the right one. In: Workshop on visibility and availability of LT resources, NODALIDA 2011, Riga, Latvia, pp 11–17, <http://hdl.handle.net/10138/29355>
- Oksanen, V., Lindén, K., & Westerlund, H. (2010) Laundry symbols and license management: Practical considerations for the distribution of LRS based on experiences from Clarin. In: Proceedings of Language Resources and Evaluation (LREC'10) Workshop on language resources: From storyboard to sustainability and LR lifecycle management, [https://www.academia.edu/18849874/Laundry\\_Symbols\\_and\\_License\\_Management\\_Practical\\_Considerations\\_for\\_the\\_Distribution\\_of\\_LRS\\_based\\_on\\_experiences\\_from\\_CLARIN](https://www.academia.edu/18849874/Laundry_Symbols_and_License_Management_Practical_Considerations_for_the_Distribution_of_LRS_based_on_experiences_from_CLARIN)
- Petrov, S., Das, D., & McDonald, R. (2012) A universal part-of-speech tagset. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey, pp 2089–2096, [http://www.lrec-conf.org/proceedings/lrec2012/pdf/274\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf)

- Piasecki, M. (2014) User-driven language technology infrastructure—the case of clarin-pl. In: T Erjavec & J Zganec Gros (eds.) Proceedings of the Ninth Language Technologies Conference—Information Society—IS 2014, Institut Jožef Stefan, Ljubljana, Slovenia, [http://nl.ijs.si/isjt14/proceedings/isjt2014\\_01.pdf](http://nl.ijs.si/isjt14/proceedings/isjt2014_01.pdf)
- Piasecki, M., Walkowiak, T., & Eder, M. (2018a). Open stylometric system WebSty: Integrated language processing, analysis and visualisation. *Computational Methods in Science and Technology*, 24(1), 43–58. <https://doi.org/10.12921/cmst.2018.0000007>
- Piasecki, M., Walkowiak, T., & Maryl, M. (2018b). Literary exploration machine a web-based application for textual scholars. In: M. Piasecki (ed.) Selected papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017. Linköping University Electronic Press, Sweden, no. 147 in Linköping Electronic Conference Proceedings, pp 128–144, <http://www.ep.liu.se/ecp/147/011/ecp17147011.pdf>
- Piasecki, M., Walkowiak, T., & Pol, M. (2017). Processing, analysing and visualising language data using solutions prepared in CLARIN-PL LTC. In: Z. Vetulani & P. Paroubek (eds.) Proceedings of human language technologies as a challenge for computer science and linguistics, Poznań, Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu, pp 292–296, <http://lfc.amu.edu.pl/book/papers/LRT2-4.pdf>
- Piperidis, S., Labropoulou, P., Deligiannis, M., & Giagkou, M. (2018). Managing public sector data for multilingual applications development. In: N. Calzolari, K. Choukri, C. Cieri, et al. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan, <https://aclanthology.org/L18-1205.pdf>
- Rehm, G. (2023). *European Language grid—a language technology platform for multilingual Europe*. <https://doi.org/10.1007/978-3-031-17258-8>
- Saulīte, B., Dargis, R., Gruzītis, N., Auzina, L., Levane-Petrova, K., Pretkalnina, P., Rituma, L., Paikens, P., Znotins, A., Stranskale, L., Pokratniece, K., Poikans, I., Barzdins, G., Skadina, I., Baklane, A., Saulespuren, V., & Ziedins, J. (2022). Latvian national corpora collection—korpuss. lv. In: Proceedings of the 13th language resources and evaluation conference (LREC), pp 5123–5129, <http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.548.pdf>
- Spektors, A., Auziņa, I., Dargis, R., Gruzītis, N., Paikens, P., Pretkalnina, L., Rituma, L., & Saulīte, B. (2016). Tēzurs.lv: the largest open lexical database for Latvian. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16), pp 2568–2571, <https://aclanthology.org/L16-1408>
- Spektors, A., Pretkalniņa, L., Grūzītis, N., Paikens, P., Rituma, L. & Saulīte, B. (2019). Tēzurs.lv 2020. <http://hdl.handle.net/20.500.12574/9>, CLARIN-LV digital library at IMCS, University of Latvia
- Straka, M., & Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared task: multilingual parsing from raw text to universal dependencies. Association for computational linguistics, Vancouver, Canada, pp 88–99, <https://doi.org/10.18653/v1/K17-3009>, <https://aclanthology.org/K17-3009>
- Van Uytvanck, D., Zinn, C., Broeder, D., Wittenburg, P., & Gardellini, M. (2010). Virtual language observatory: The portal to the language resources and technology universe. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association, pp 900–903, [http://www.lrec-conf.org/proceedings/lrec2010/pdf/273\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2010/pdf/273_Paper.pdf)
- Walkowiak, T. (2018). Language Processing Modelling Notation-Orchestration of NLP Microservices. *Advances in Intelligent Systems and Computing*, 582, 464–473. [https://doi.org/10.1007/978-3-319-59415-6\\_44](https://doi.org/10.1007/978-3-319-59415-6_44)
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S. A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3(1), 1–9. <https://doi.org/10.1038/sdata.2016.18>

- Yimam, S. M., Gurevych, I., de Castilho, R. E., & Biemann, C. (2013) Webanno: A flexible, web-based and visually supported system for distributed annotations. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL'13). Association for Computational Linguistics, pp 1–6, <http://tubiblio.ulb.tu-darmstadt.de/98019/>
- Zinn, C. (2018a). A bridge from EUDAT's B2DROP cloud service to CLARIN's Language Resource Switchboard. In: Selected papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017, Linköping University Electronic Press, no. 147 in Linköping Electronic Conference Proceedings, pp 36–45, [https://www.clarin.eu/sites/default/files/Zinn-CLARIN2017\\_paper\\_17.pdf](https://www.clarin.eu/sites/default/files/Zinn-CLARIN2017_paper_17.pdf)
- Zinn, C. (2018). The language resource Switchboard. *Computational Linguistics*, 44, 1–13.
- Zinn, C., & Campbell, B. (2022). WebLicht-Batch—a web-based interface for batch processing large input with the WebLicht workflow engine. In: Proceedings of the CLARIN Annual Conference, Prague, [https://www.clarin.eu/sites/default/files/CLARIN2022\\_P\\_2.1.2\\_ZinnCampbell.pdf](https://www.clarin.eu/sites/default/files/CLARIN2022_P_2.1.2_ZinnCampbell.pdf)
- Znotins, A., & Cirule, E. (2018). NLP-pipe: Latvian NLP tool pipeline. *Human Language Technologies—The Baltic Perspective*, 307, 183–189. <https://doi.org/10.3233/978-1-61499-912-6-183>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

A. Branco<sup>1</sup>  · M. Eskevich<sup>2</sup> · F. Frontini<sup>3</sup> · J. Hajič<sup>4</sup> · E. Hinrichs<sup>5</sup> · F. de Jong<sup>2</sup> · P. Kamocki<sup>6</sup> · A. König<sup>2</sup> · K. Lindén<sup>7</sup> · C. Navarretta<sup>8</sup> · M. Piasecki<sup>9</sup> · S. Piperidis<sup>10</sup> · O. Pitkänen<sup>11</sup> · K. Simov<sup>12</sup> · I. Skadiņa<sup>13</sup> · T. Trippel<sup>14</sup> · A. Witt<sup>15</sup> · C. Zinn<sup>5</sup>

✉ A. Witt  
witt@ids-mannheim.de

A. Branco  
antonio.branco@di.fc.ul.pt

M. Eskevich  
maria@clarin.eu

F. Frontini  
francesca.frontini@ilc.cnr.it

J. Hajič  
hajic@ufal.mff.cuni.cz

E. Hinrichs  
erhard.hinrichs@uni-tuebingen.de

F. de Jong  
f.m.g.dejong@uu.nl

P. Kamocki  
kamocki@ids-mannheim.de

A. König  
alex@clarin.eu

K. Lindén  
kristen.linden@helsinki.fi

C. Navarretta  
costanza@hum.ku.dk

M. Piasecki  
maciej.piasecki@pwr.edu.pl

S. Piperidis  
spip@athenarc.gr

O. Pitkänen  
olli.pitkanen@1001lakes.com

K. Simov  
kivs@bultreebank.org

I. Skadiņa  
inguna.skadina@lumii.lv

T. Trippel  
thorsten.trippel@uni-tuebingen.de

C. Zinn  
claus.zinn@uni-tuebingen.de

- <sup>1</sup> University of Lisbon, Lisbon, Portugal
- <sup>2</sup> CLARIN ERIC, Utrecht, Netherlands
- <sup>3</sup> CNR-ILC and CLARIN ERIC, Pisa, Italy
- <sup>4</sup> Charles University, Prague, Czechia
- <sup>5</sup> University of Tübingen, Tübingen, Germany
- <sup>6</sup> Leibniz Inst. for the German Lang., Mannheim, Germany
- <sup>7</sup> University of Helsinki, Helsinki, Finland
- <sup>8</sup> University of Copenhagen, København, Denmark
- <sup>9</sup> Wrocław Univ. of Sci. and Techn., Wrocław, Poland
- <sup>10</sup> Athena Research Center, Marousi, Greece
- <sup>11</sup> 1001 Lakes, Helsinki, Finland
- <sup>12</sup> Institute of Information and Communication technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria
- <sup>13</sup> Institute of Mathematics and Computer Science, University of Latvia, Rīga, Latvia
- <sup>14</sup> University of Tübingen and Leibniz Institute for the German Language, Tübingen, Germany
- <sup>15</sup> Leibniz Institute for the German Language, Mannheim, Germany