information

Article

Exploring Reusability and Reproducibility for a Research Infrastructure for L1 and L2 Learner Corpora

Alexander König [1,*], Jennifer-Carmen Frey [2] and Egon W. Stemle [2]

[1] CLARIN ERIC, 3512 BS Utrecht, The Netherlands
[2] Institute for Applied Linguistics, Eurac Research, 39100 Bolzano, Italy; jennifercarmen.frey@eurac.edu (J.-C.F.); egon.stemle@eurac.edu (E.W.S.)
* Correspondence: alex@clarin.eu

Abstract: Up until today research in various educational and linguistic domains such as learner corpus research, writing research, or second language acquisition has produced a substantial amount of research data in the form of L1 and L2 learner corpora. However, the multitude of individual solutions combined with domain-inherent obstacles in data sharing have so far hampered comparability, reusability and reproducibility of data and research results. In this article, we present work in creating a digital infrastructure for L1 and L2 learner corpora and populating it with data collected in the past. We embed our infrastructure efforts in the broader field of infrastructures for scientific research, drawing from technical solutions and frameworks from research data management, among which the FAIR guiding principles for data stewardship. We share our experiences from integrating some L1 and L2 learner corpora from concluded projects into the infrastructure while trying to ensure compliance with the FAIR principles and the standards we established for reproducibility, discussing how far research data that has been collected in the past can be made comparable, reusable and reproducible. Our results show that some basic needs for providing comparable and reusable data are covered by existing general infrastructure solutions and can be exploited for domain-specific infrastructures such as the one presented in this article. Other aspects need genuinely domain-driven approaches. The solutions found for the corpora in the presented infrastructure can only be a preliminary attempt, and further community involvement would be needed to provide templates and models acknowledged and promoted by the community. Furthermore, forward-looking data management would be needed starting from the beginning of new corpus creation projects to ensure that all requirements for FAIR data can be met.

Keywords: learner corpus research; research infrastructures

Citation: König, A.; Frey, J.-C.; Stemle, E.W. Exploring Reusability and Reproducibility for a Research Infrastructure for L1 and L2 Learner Corpora. *Information* **2021**, *12*, 199. https://dx.doi.org/10.3390/info12050199

Academic Editor: Dragan Ivanovic

Received: 26 February 2021
Accepted: 28 April 2021
Published: 30 April 2021

1. Introduction

Various fields in educational research and applied linguistics work with language data produced by writers or speakers who are still acquiring language competence in the language or language variety they use to express themselves. Most evidently, this regards language produced by non-native speakers, i.e., language learners, where fields such as second language acquisition, learner corpus research, and educational research in language learning and teaching have a long-standing tradition. In contrast, newer fields such as computer-assisted language learning have become very prominent in recent times due to the technological advances of recent decades. On top of this, language produced by novice, not fully proficient native speakers has come into focus due to increasing amounts of writing and language assessment research and research on literacy development.

The data used in these domains are usually large electronic collections, i.e., corpora of texts or utterances, combined with relevant metadata on the authors and the context of the production. In many cases, the data also include annotations on the texts that highlight certain phenomena of interest. Corpora depicting language production by non-native

speakers in a second (or third) language (L2) are typically called *learner corpora*, with an entire research field, learner corpus research [1], specializing on the use of these resources to investigate the dynamics and outcomes of language learning processes on empirical data (cf. [2]). In its three decades of existence, the learner corpus research community has created a remarkable number of learner corpora and tools used in analyses both by the original data collectors and—when data or source code had been made available—by other interested researchers from the field. Available tools tailored explicitly for the use in learner corpus research, such as EXMARaLDA [3], the Sketch Engine for Language Learning (SkELL) [4], ANNIS [5], TEITOK [6], Transc&Anno [7] and Korp searches in Second Language data [8], are usually shared as downloadable, executable applications or through source-code sharing platforms and thus follow a long tradition of software sharing practices that has established solutions for aspects such as licensing. Sharing data, however, is not equally streamlined yet, and thus publishers of L2 learner corpora have found various, individual solutions on how to make data available to others (see, for example, the L2 corpora listed by the Center for English Corpus Linguistics at the University of Louvain [9] or in the CLARIN resource family for L2 corpora [10] (https://clarin.eu/resource-families/L2-corpora, accessed on 22 April 2021)). This is equally true for L1 corpora, where research has also led to a substantial number of resources that have been created so far (see, e.g., the LOCNESS corpus (https://www.learnercorpusassociation.org/resources/tools/locness-corpus/, accessed on 22 April 2021) or the Litkey corpus (https://www.linguistics.rub.de/litkeycorpus/, accessed on 22 April 2021)).

Although there is not much overlap between researchers analyzing L2 learner corpora and those researching student essays, early academic writing or other non-professional productions by native speakers (henceforward called L1 corpora.), both face very similar issues when collecting, preparing, using and sharing data and the tools they use are often the same (see [11] for an extensive argumentation on this). We, therefore, argue that they can be treated similarly from a data management and data sharing point of view. Both L1 and L2 corpora need to be systematic and purposefully sampled according to the envisioned research (for L2 learner corpora, see [12] or [13], but this is a general characteristic of corpora [14,15]). They are usually laboriously collected, for example, through the interaction with language learning institutions, and apart from the need for arduous acquisition and management of data usage consents, the texts are frequently elicited through pen and paper tests or recordings of spoken utterances that need (manual) transcription before any other processing step. Moreover, before being used, corpora often undergo manual or semi-manual annotation processes. Additional information is added to the plain text material to provide researchers with (searchable) frequency information on specific linguistic phenomena of interest. Only in this way are the now significantly enriched datasets of value to the researcher and the wider research community.

However, more than just differing between L1 and L2 data, the corpus designs and corpus sharing solutions found so far differ on a more fundamental level from resource to resource. There are various individual solutions for collecting, pre-processing and annotating L1 and L2 corpora that exceed the mere distinction of the research domain but go down to individual research projects. Similarly, the solutions for sharing data with the research community are manifold and although, of course, not all resources are shared with the broader scientific community but are only available for the members of a particular research institution or are even proprietary to a single researcher, even the ones that are available to the general public show relatively low interoperability and comparability [16]. For the progress of the field, more methodological rigor is needed [17,18], including increased transparency, the use of more standardized and well-documented methods, and the reproducibility of research results, which is not necessarily given if data is not made available in a useful way. All this has set the grounds for the recent attention to standardization of L1 and L2 corpora, both between and among themselves [19,20] and the interest in creating digital infrastructures that make such corpora and corresponding tools available to the academic public (e.g., ref. [21,22]).

Recent advances in infrastructures for scientific research further encourage this development. The strategic turn towards establishing international and partly cross-discipline infrastructures has been promoted through the help of funding bodies for scientific research. An important waypoint in this development was the foundation of the Research Data Alliance (RDA) [23] in 2013, a global body to discuss issues of standardization and interoperability internationally on a high level. In Europe, the European Commission has made this one of the main points of its funding strategies, first with the establishment of European Research Infrastructure Consortia (ERICs) (https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures/eric_en, accessed on 22 April 2021) [24] in 2009 and then with the more broad initiative of the European Open Science Cloud (EOSC) (https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/european-open-science-cloud-eosc_en, accessed on 22 April 2021) [25] Building the European Open Science Cloud as an overarching research infrastructure with the aim to support all scientific research within Europe (and beyond) has started in 2015 and is currently being worked on in multiple large-scale cross-discipline multi-national European-funded projects such as SSHOC (Social Sciences and Humanities Open Cloud) (https://www.sshopencloud.eu/, accessed on 22 April 2021) [26], TRIPLE (Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration) (https://www.gotriple.eu, accessed on 22 April 2021) [27] and many others. The goal of all these efforts is to establish long-term sustainable infrastructures for scientific research that help researchers in their work by providing them with tools, training, data storage and computing power. The focus lies on Open Science and especially the accordance with the FAIR principles [28].

The European Open Science Cloud (EOSC) is an initiative launched by the European Commission in 2015 to promote open science within the European research community. The idea is to build sustainable infrastructures for scientific research that help researchers search for and access existing data—either from their own or from neighboring disciplines—and software that they can use to process their data. Building and enhancing the EOSC is currently one of the key strategies of European research funding. This is done mostly by interconnecting all the existing tools and infrastructures that already exist to ensure that users can easily find the tools and data they are looking for and find them in a format that is useful to them (https://eosc-portal.eu/, accessed on 22 April 2021).

SSHOC, the Social Sciences and Humanities Open Cloud, is one of the projects that the EU currently funds as part of the wider EOSC strategy. It focuses on the social sciences and humanities. One of the key features of the project is the development of an "SSH Open Marketplace" (https://marketplace.sshopencloud.eu/, accessed on 22 April 2021) [29] that will collect information about software, workflows, datasets and scientific papers and make them easily searchable while also interlinking them with each other. The platform will make it possible to look, for example, at a software tool and receive information on various papers mentioning this tool or to read about a workflow that has all necessary tools and services for each of the steps linked to it.

Within the EOSC infrastructure and SSHOC, the CLARIN community (https://www.clarin.eu, accessed on 22 April 2021) [30] focuses on language data and is therefore the general infrastructure community most relevant to research data infrastructures for L1 and L2 corpora. CLARIN—the Common Language Resources and Technology Infrastructure—is set up as an ERIC comprising 21 member countries and three observers. This means that partners in all these countries are contributing to the European infrastructure, for example, by running a center providing language data and tools.

As can be seen, the creation of digital research infrastructures to help researchers with the discoverability and availability of data, software, and methods is currently a very prominent topic. This aligns with the needs identified in the L1 and L2 learner corpus research community.

In this article, we describe the design of an infrastructure for L1 and L2 corpus research that has been integrated into the wider CLARIN infrastructure and aims to provide reusable

language resources, both for L1 and L2 corpora and related processing tools. The two main goals that led our activities in creating the infrastructure were to

1. provide harmonized L1 and L2 corpora with a clear focus on reusability of the data and
2. ensure reproducible research results by integrating research workflows in L1 and L2 corpus research.

For setting up the technical basis of the infrastructure, we relied on established solutions from the general research infrastructure community that apply to our domain, as described in Section 3. For testing whether the infrastructure allows us to provide harmonized L1 and L2 corpora that are reusable and ensure reproducible research results, we integrated the L1 and L2 corpus resources collected at a European research center into the infrastructure. We referred to the FAIR guidelines for data stewardship [28] in order to prepare data and metadata that are Findable, Accessible, Interoperable and Reusable, considering that [28] pointed out that all four aspects need to be taken into account to provide reusable data. Finally, for tackling reproducibility of research results, we identified two main prerequisites. First, proper versioning of the data so that it is clear which analysis had been done on which version of the data. Second, reproducibility of processing pipelines, for which we investigated the use of containerized workflows, thus encapsulating all the tools necessary to recreate certain processing steps. However, this part of the infrastructure is currently only theoretical, and only preliminary experiments have been carried out.

In the following, we present our activities. Section 2 introduces the research institute and their L1 and L2 corpora that were to be integrated into the future research infrastructure for L1 and L2 corpora. It presents domain-specific options for making data available and addresses issues regarding the FAIR-compatible provision of L1 and L2 corpora in general and of the data to be integrated. In Section 3, we present the basic design, technical implementation and supporting activities of the newly created infrastructure for L1 and L2 corpora (see Section 3.1). We share considerations on how to achieve more reproducibility in the data-processing pipeline (Section 3.2) and explain in detail our solutions integrating the corpus resources in the infrastructure while complying to the FAIR principles. Finally, Section 4 discusses our results while pointing out difficulties that can only be tackled in community-driven future work.

2. Materials and Methods

2.1. Data

Over the last 15 years, the Institute for Applied Linguistics at Eurac Research (IAL) in Bolzano, Italy, has conducted or partnered in five research projects concerned with the empirical analysis of texts written by students in their second or first language. The resources originating from these projects are manifold and comprise L1 and L2 learner corpora in four languages (although focusing on Italian and German as the main languages of the local territory) and from various language backgrounds (German or Italian respectively, bilingual German and Italian writers, writers with other language backgrounds due to recent migration of their family). Many of the resources are, however, related to each other and allow for comparison between different text genres, between L1 or L2 writings, or between L2 writers of various L1 backgrounds. This makes it relevant to provide them in a unified and standardized manner which makes such comparisons among different resources possible. Furthermore, the data collection workflow was similar for all resources, allowing the retrieval of best practices and lessons learned from the experiences. All corpora were elicited in educational settings, were produced in handwritten format and digitized and processed later through manual transcription. None of the corpora contained multi-modal or spoken language. Table 1 gives a short overview of the resources created so far. Moreover, the institute already plans further projects in these fields, collecting further data and enriching their portfolio. One of which is collecting and analyzing an Italian L1 corpus as a complementary resource to the KOKO corpus mentioned below.

The data that have been used for corpus linguistic analyses in the past offer, on the one hand, valuable resources to share with the community; on the other hand, they represent a rich test bed for standardization approaches relevant to the community as well as for experimenting with best practices on how to share L1 and L2 corpora in a sustainable way allowing for later comparability, reusability and reproducibility.

Table 1. Overview of corpora to integrate into the infrastructure

Corpus	Size in Texts	Text Language	Data Collection Time
Kolipsi-1_L2	ca. 2 500	German, Italian (L2)	2007
Kolipsi-1_L1	ca. 500	German, Italian (L1)	2010
Kolipsi-2	ca. 2 500	German, Italian (L2)	2014
KoKo	ca. 1 500	German (L1)	2011
LEONIDE	ca. 2 500	German, Italian, English (L1, L2, L3)	2015-2018
Merlin	ca. 2 300	German, Italian, Czech (L2)	2012

2.1.1. Kolipsi-1 (L2)

The Kolipsi-1 (L2) corpus (http://hdl.handle.net/20.500.12124/26, accessed on 22 April 2021) is a typical L2 learner corpus. It contains German and Italian L2 productions. All the writers were upper secondary school students whose native language was, for the most cases, the respective other language, that is, German when the text was written in Italian and vice versa. Besides, there was also a small number of students with migratory background having neither German nor Italian as L1 but being educated in one of the two languages in school. Students performed two writing tasks that consisted of an e-mail to a friend describing a picture story (narrative text genre) and writing a letter to a friend on holiday planning (argumentative text genre). All data were collected in the school year 2007/2008 [31]. The handwritten texts were afterwards manually transcribed and annotated for surface features such as student corrections or graphical elements as well as orthographic errors, including a normalized form (i.e., target hypothesis) and automatically for sentence splitting, tokenization, lemmatization and part-of-speech tagging. Metadata from the writers was elicited using a questionnaire and contained various items relevant for educational research such as school type, gender, place of origin, socio-economic background and L1 of the learners.

2.1.2. Kolipsi-1 (L1)

Kolipsi-1 (L1) (http://hdl.handle.net/20.500.12124/29, accessed on 22 April 2021) is a small reference corpus built for comparison of L2 and L1 writings of the same tasks in the Kolipsi 1 project. The data was collected from Italian and German L1 speakers in Italy and Germany in 2010. The elicitation method and design were the same as in Kolipsi-1 L2.

2.1.3. Kolipsi-2

In 2014 a repetition of the Kolipsi-1 study was initiated that aimed to compare the data retrieved in 2007 with new data from the local territory to observe differences in the average language competence of the population in their second language at the time of high school graduation seven years after the project was first conducted [32]. The data coming from this replication study led to the Kolipsi-2 corpus (http://hdl.handle.net/20.500.12124/30, accessed on 22 April 2021) that was built in analogy to the Kolipsi-1 corpora. The design and setup of the study, as well as the processing of the data, stayed the same. However, one of the writing tasks (argumentative text) changed slightly.

2.1.4. KoKo

KoKo (http://hdl.handle.net/20.500.12124/12, accessed on 22 April 2021) [11,33] is a German L1 corpus of argumentative student essays collected from different German-speaking regions to illustrate pluricentric language varieties. A total of 1503 texts was elicited during school classes in 2011. Students were 17–19 years old. The handwritten

texts were afterwards transcribed and annotated automatically for sentence splitting, tokenization, lemmatization, and part-of-speech tagging; and the texts were manually annotated for surface features such as student corrections or graphical elements as well as orthographic errors including the assignment of a normalized form (i.e., target hypothesis). Further error annotations done on the whole corpus regarded punctuation errors, while grammar errors and lexical misuse were annotated only for a subset of 597 and 980 texts, respectively. Metadata from the writers contained age, gender, school type, German grade, region of residence and others. Further metadata on aspects of text quality is available for a subset of 569 texts.

2.1.5. Merlin

Merlin (http://hdl.handle.net/20.500.12124/6, accessed on 22 April 2021) [34] is a trilingual learner corpus illustrating European reference levels. The corpus contains 2290 learner texts produced in standardized language certifications covering Common European Framework of Reference for Languages (CEFR) levels A1–C1 of L2 German, Italian or Czech with differing L1 backgrounds of the writers. Writing tasks differed for and within the levels; however, they typically consisted of replying to a prompt by writing a letter to a friend or business. The metadata available for the writers contains information on age, gender, and first language. Additionally, information on the CEFR test level, the test institution, test task and target language is given. The texts were extracted from the original tests, rated according to a CEFR level compliant rating grid and annotated manually with explicit target hypothesis and annotations for linguistic phenomena and errors on orthography, grammar, vocabulary, coherence/cohesion, sociolinguistic appropriateness, pragmatics and others. Automatic linguistic annotation included sentence splitting, tokenization, lemmatization, part-of-speech tagging and syntactic parsing.

2.1.6. LEONIDE

The data collected during the project One School many Languages (http://hdl.handle.net/20.500.12124/25, accessed on 22 April 2021) represents a longitudinal corpus of German, Italian and English texts composed by the same middle school students during three subsequent school years. Students were mainly German or Italian native speakers, although more complex language biographies were found in the South Tyrolean school sample (cf. [35]). The dataset that was designed to allow the analysis of multilingual competences contains 1265 picture stories and 1265 opinion texts handwritten by the students. The texts were subsequently transcribed and annotated manually for surface features such as student corrections or graphical elements as well as orthographic errors, including a normalized form (i.e., target hypothesis) and automatically for sentence splitting, tokenization, lemmatization and part-of-speech tagging.

2.2. Making L1 and L2 Corpus Data Available to the Research Community

There are mainly two ways to provide L1 and L2 corpora to the public: (a) the provision of the corpus as a searchable resource (i.e., the provision of a query interface with a so-called concordancer to search word forms, structures or other linguistic phenomena in the corpus) or (b) the provision of a corpus as open and reusable research data (e.g., through download options). Examples of well-known learner corpora that come with their own query interfaces are the International Corpus of Learner English (ICLE) [36] or the Norsk andrespråkskorpus (ASK) [37]. Corpora available as download to use freely without any technical restrictions are, for example, the Czech as a Second Language corpus (CzeSL) [38] or the ETS Corpus of Non-Native Written English [39].

Although both ways to provide L1 and L2 learner corpora are generally independent of each other, the first poses the necessity to set up and sustain a (custom-made) query interface while posing challenges on reproducibility and comparability of research results. This can be mitigated by providing a lot of corpora through the same query interface as is being done by CLARIN.SI (https://www.clarin.si, accessed on 22 April 2021) or in the

infrastructure that is presented in this paper. The latter way, on the other hand, comprises additional copyright and licensing concerns because the full data is accessible, as well as challenges regarding how and where to provide the data to ensure its reusability.

In general, the provision of L1 and L2 learner corpora through query interfaces is often preferred by corpus creators since it is less permissive than the provision of downloadable corpus data. It avoids copyright issues as search interfaces often prohibit visualization or download of full texts and prevents the alteration of the data. Also, the typical corpus user might appreciate the availability of a ready-made search interface without needing to deal with different file formats and tools to perform classical corpus linguistic analyses such as the visualization of key-word-in-context overviews or the extraction of frequency lists. However, each query interface also restricts the use of the resource to the functions and use cases implemented in it, thereby limiting its use to those who are (a) familiar and (b) satisfied with the data, methods, and possibilities provided. Furthermore, setting up and sustaining a (custom-made) query interface for a corpus needs a lot of background knowledge, technical skills and resources and might not be feasible for smaller research projects in L1 and L2 learner corpus research. The exclusive provision of concordancers also comes with caveats regarding the comparability and reproducibility of research results. Comparisons of research results across corpora provided in different platforms are rarely possible due to the difference in functionalities and used configurations (e.g., for how they calculate corpus statistics). Corpus updates or modifications in the configurations and functions might prohibit the reproducibility of research results unless older versions stay available for reference. A full corpus download, on the other hand, would often allow more usage scenarios reaching other potential target groups that, for example, work on different research questions or with different methods (e.g., NLP research, more complex quantitative analysis, or simply the upload of the data in the query interface of choice).

2.3. Ensuring the Provision of FAIR Data

Although issues of comparability and reproducibility can be addressed with the provision of open research data allowing consequently also for more transparency in the field, making the provided data interoperable and reusable—and thus valuable for the community—needs directed steps. The FAIR guiding principles for data stewardship provide a framework for tackling this aim. By referring to the guidelines, best practices for data sharing can be defined for individual communities where basic requirements for providing findable, accessible, interoperable and reusable research data should be accounted for [28,40].

In the following, we will look at each of the FAIR principles in detail and discuss them in the context of current practices within the field of L1 and L2 learner corpus research.

2.3.1. Findability

The first and most basic step in making research data FAIR, and thus ultimately reusable, is to make them findable. [28] define a clear set of recommendations for ensuring the Findability of research data. Findable data (and their metadata) are assigned a globally unique persistent identifier (F1), they are described with rich metadata (F2), which clearly and explicitly includes the identifier of the data they describe (F3) and both data and metadata are registered or indexed in a searchable resource (F4).

However, when [41] conducted a survey on existing L2 learner corpora regarding their Findability in 2018, they noticed that it is one of the major issues for this type of language resources (next to insufficient availability of metadata and documentation). Out of the 180 L2 learner corpora they investigated (a subset of the UCL list mentioned above), only 31 were indexed in the Virtual Language Observatory (VLO) (https://vlo.clarin.eu/, accessed on 22 April 2021), a well-known search engine for language resources in linguistics that allows searching on various metadata fields [42]. On the one hand, the lacking Findability of the language resources is certainly due to the many corpora that are produced in short-lived projects and mentioned only in the derived research articles, without ever

being deposited in research data repositories, assigned a persistent identifier or at least documented in machine-readable form (see also [41]). Nevertheless, even for deposited data, missing metadata and the lack of the technical prerequisites to search for learner corpus-specific categories in the search interface meant that Findability was limited.

2.3.2. Accessibility

In terms of Accessibility, the guiding principles define that metadata and data should be retrievable by their identifier using a standardized communications protocol (A1). The protocol should be open, free and universally implementable (A1.1) and should allow for an authentication and authorization procedure, where necessary (A1.2). Furthermore, metadata should be accessible, even when the data are no longer available (A2). However, according to the evaluation of [41], only 23 out of the 31 L2 learner corpora that they identified as findable were also available for download or querying and thus accessible in some way. In many cases, this is due to legal constraints of the data, which are one of the key obstacles in providing such type of data to the public. Especially missing or too narrow user consents, but also the need for manually conducted anonymization or pseudonymization constrains the provision of data (see, e.g., ref. [43]). As legal requirements for publishing the corpus to a broader audience are most often country-dependent and contradicting [44], this is a serious issue for L1 and L2 corpora that often contain rather sensitive data of minors (see also [20]). However, while it might not be possible to provide data itself, the FAIR principles require that the metadata should be made available in a standardized accessible way so that researchers are able to learn from previous corpus creation projects, compare their research and designs and potentially draw from existing knowledge while making resources and studies more comparable (cf. also [41]).

2.3.3. Interoperability

Interoperability of L1 and L2 corpora is another important step towards reusable corpus data and sets the foundations for comparing and integrating or even aggregating data and approaches. For this, the FAIR principles recommend using a formal, accessible, shared and broadly applicable language for knowledge representation for data and metadata (I1), to use vocabularies for data and metadata that themselves follow the FAIR principles (I2) and to include qualified references to other data or metadata, if relevant (I3). The Interoperability of L1 and L2 learner corpora has received increasing attention in recent years as researchers have identified the need for more consensus and international collaboration to make learner corpus research more comparable and as such interoperable (e.g., ref. [20,44]).

[44], for example, state "there is a need to make sure that L2 corpora have comparable error taxonomies (i.e., mark-up for deviations in orthography, tense, etc.), associated metadata variables (e.g., age, gender, task, etc.), file formats (e.g., JSON, XML), corpus design (e.g., L1 grouping), etc." And furthermore, when looking at the formats used for knowledge representation (i.e., structural interoperability), no formally established standards have been defined for the domain of learner corpora.

For the language data itself, many researchers use the XML format of the Text Encoding Initiative (TEI) (https://tei-c.org/, accessed on 22 April 2021) ([20] therefore names it the "de-facto" standard for learner corpora) in combination with the IMS corpus workbench (http://cwb.sourceforge.net/, accessed on 22 April 2021) for corpus query or PAULA/XML (http://www.sfb632.uni-potsdam.de/en/paula.html, accessed on 22 April 2021) in combination with ANNIS (https://corpus-tools.org/annis/, accessed on 22 April 2021) for corpus query [44]. As can be seen in those two examples, for corpus research, not only the knowledge representation in the data files but also the choice of query interfaces can limit or enhance the Interoperability of resources. For the corpora represented in TEI XML, the metadata is often defined in the so-called TEI Header, which offers a set of useful metadata categories for knowledge representation [44].

Although this approach is convenient and coherent in terms of data provision, combining metadata and data in one file poses the risk that metadata will be unavailable when the data is not available any longer, which generally goes against the recommendation for FAIR research data.

In terms of conceptual interoperability, ref. [44] and [45] state that "there is increasing convergence, however, on the need of one relatively stable set of recommendable or obligatory metadata for learner characteristics, and on another set for corpus information" and that the use of previously established metadata sets by various further projects, all using TEI Header could lead to "what could become a generally accepted extension to the TEI standard". However, there are no FAIR vocabularies for the learner corpus domain yet. Ongoing SSHOC activities (see https://www.sshopencloud.eu/news/workshop-notes-sshoc-requirements-vocabularies-and-vocabulary-management-platforms, accessed on 22 April 2021) aim to solve the technical issues in making vocabularies findable, accessible, interoperable and reusable once they are established (e.g., assigning them persistent identifiers, making them available via standardized protocols as compared to privately curated and distributed lists, presenting them in a formal language for knowledge representation and describing them thoroughly). However, creating the vocabularies depends on the domain itself. Finally, the FAIR guidelines recommend making qualified references to other relevant data and metadata, which would make a comparison between similar or even related data easier. However, few learner corpora are described next to each other, with comparable information on the same platforms, or can be queried with the same query interface. Usually, corpora are described on their own individual web pages or come with their own search interfaces that very rarely give direct links to other data.

2.3.4. Reusability

Finally, to reach the ultimate goal of reusable data, ref. [28] state that metadata and data should be well-described enough to be replicable and combinable in different settings. For this, metadata and data should be released with a clear and accessible data usage license (R1.1), associated with detailed provenance (R1.2) and compliant to domain-relevant community standards (R1.3). For L1 and L2 learner corpora, this means that extensive documentation is needed for making the data reusable that includes in particular domain-specific information next to detailed information about the purpose and circumstances of data collection. This comprises information on:

- the background of the writers (e.g., L1, age, proficiency, other L2s)
- the writing task itself (target language, genre, writing prompts etc.)

Without this information, it is impossible to judge whether a corpus is a suitable resource for one's own research questions [13].

However, it has been observed that the metadata provided for learner corpora so far often strongly depends on the focus of the underlying research and thus "shows substantial variation" [44]. Ref. [41] and [46] point out that crucial metadata information is frequently missing, proposing the creation of a standardized core metadata set that can be re-used for future corpus collection projects. This core metadata set should contain the items mentioned above, but also information on the provenance of the data (e.g., authors, responsible people for data collection, processing and annotation, time and place of data collections) and licensing information. Both of which are currently often neglected (see also [41] who only found licensing information for 28 out of 31 findable corpora).

3. Results

3.1. Base Components and Implementation of the Infrastructure

To achieve reproducibility and reusability while reaching a mostly non-technical target audience, we decided to provide both downloadable resources as well as a harmonized web-based search interface for accessing the corpora of the infrastructure. Additionally, we provide a central access point in the form of a web portal that gives further information and

documentation on the corpora, the research related to those corpora and the underlying infrastructure.

In the following, we describe the individual components created for this aim.

3.1.1. Porta—A Central Access Point for SSH Researchers

Porta (https://www.porta.eurac.edu/, accessed on 22 April 2021) is a platform directed at SSH researchers that should provide a central access point to the L1 and L2 learner corpora provided by the infrastructure, giving access to documentation, data downloads and search interfaces for all resources. It is directed at a less-technical audience, maintained as a WordPress installation that allows for rights management and back-end access to various researchers, engaged with the data creation and provides unified templates for integrating and introducing new corpora.

Each corpus pages consists of the following three main components (while allowing the addition of other components if needed).

- Human-readable and unified documentation
  The documentation presents the
  - background of the corpus and its corpus design
  - statistics on the corpus such as number of texts and tokens, number of writers, represented languages and language backgrounds of the writers
  - transcription and annotations guidelines,
  - annotation schemes used and description of corpus creation procedures
- Links to the FAIR data resources in a long-term archive (see Section 3.1.2 for details)
- A unified search interface to query the corpora directly(see Section 3.1.3 for details).

3.1.2. CLARIN-DSpace

CLARIN-DSpace is an open-source repository software that is ideally suited for running a CLARIN Center and is developed by multiple CLARIN Centers across Europe under the lead of the developers at LINDAT/CLARIAH-CZ, the national CLARIN consortium in the Czech Republic.

CLARIN-DSpace is meant to cover all the necessary prerequisites for running a CLARIN B Center [47] (https://www.clarin.eu/content/assessment-procedure, accessed on 22 April 2021). It is based on a widely used open-source software (https://duraspace.org/dspace/, accessed on 22 April 2021) and tries to keep the adaptions and additions to a minimum to reduce the amount of additional maintenance as much as possible. The code for CLARIN-DSpace is maintained in a public git repository (https://github.com/ufal/clarin-dspace/, accessed on 22 April 2021), and the repository software is currently run by 14 CLARIN Centers across Europe and therefore represents a large share in the relatively varied "repository landscape" of CLARIN [48].

The central component of CLARIN-DSpace is its function as a repository for research data. It provides an easy way to deposit such data while ensuring that all the necessary information (metadata) is provided by the depositor. Each deposit will get a persistent identifier (based on the handle system) and automatically be registered with several domain-relevant search engines, most importantly the CLARIN Virtual Language Observatory (VLO) (https://vlo.clarin.eu/, accessed on 22 April 2021) [42] and the search at the Open Language Archives Community (OLAC) (http://search.language-archives.org/, accessed on 22 April 2021). This is done using the OAI-PMH standard (https://www.openarchives.org/pmh/, accessed on 22 April 2021) for metadata harvesting to ensure wide interoperability.

Another important part of the software is its integration into the CLARIN federated identity (https://www.clarin.eu/content/federated-identity, accessed on 22 April 2021) using the Shibboleth AAI. This way, users can log in to the system using their regular academic accounts from their university. The federation is based on eduGAIN and therefore can be used as a soft guarantee to ensure that users come from an academic institution. Additionally, all deposits are assigned a clear and transparent license so that it will always

be clear to a user what they can do with a certain deposit and what uses are not permitted, e.g., commercial uses.

The CLARIN-DSpace instance running at the Eurac Research CLARIN Center (https://clarin.eurac.edu, accessed on 22 April 2021) is adapted to fit with the overall design of the ERCC using the customization features provided by the software. Additionally, it runs as a Docker container within a Kubernetes infrastructure. This makes deployment and maintenance of the running software easier, and at the same time, this adaption is itself provided openly on gitlab (https://gitlab.inf.unibz.it/commul/docker/clarin-dspace, accessed on 22 April 2021) to help other centers who are planning to use it. Dockerization means that it is possible to host the software with a commercial company (e.g., Google (https://cloud.google.com/kubernetes-engine/, accessed on 22 April 2021) or Amazon (https://aws.amazon.com/kubernetes/, accessed on 22 April 2021)) and to reduce the technical overhead that must be maintained at the actual center to a minimum.

3.1.3. ANNIS

ANNIS is an open-source, versatile web-browser-based search and visualization architecture for linguistic corpora with complex multi-level annotations. ANNIS addresses the need to visualize annotations covering different linguistic levels, such as syntax, semantics, morphology, but also the need to cover the same level multiple times in different ways. It also provides means to build complex queries, for which there exists an optional graphical user interface. ANNIS is the corpus query tool where all the corpora from Porta are readily available. Corpora that are freely available in CLARIN-DSpace are configured for immediate use, without login. Other corpora require a user account, which, at the moment, must be requested by mail. In the future, users will be able to authenticate using the CLARIN federated identity and automatically gain access to appropriately licensed resources.

3.2. Integrating L1 and L2 Learner Corpus Research Workflows for Reproducibility

In research on L1 and L2 learner corpora, transparency and reproducibility of research results play an important role in advancing the field [17,18]. Apart from making research results reproducible, there is also another, more practical need for integrating workflows and providing research infrastructures. For long-lasting projects (e.g., longitudinal studies) which require the processing of new data, it is critical to have and maintain access to NLP tools in their original working state to reconstruct previous processing pipelines and always process data in an identical manner [22,49].

Compiling language data for research is often an intricate task. On the technical side, this can start as basic as the need to digitize data, which had not been born digitally and may continue with both automatic and manual processing steps. Automatic processing steps towards linguistically enriched raw text typically include tokenization, lemmatization and part-of-speech tagging, sometimes also named entity recognition and syntactic parsing. Manual processing usually consists of adding relevant information as annotations. In L1 and L2 learner corpus research, this often involves a normalized form (the so-called target hypothesis) and various encoding of errors made by the author.

Since the unification of all relevant NLP tools for a project into a single processing framework is the exception rather than the rule, this inevitably leads to a multitude of individual solutions with individual installation procedures, different development life cycles with their maintenance and update schedules, varying technical support, and so on. Furthermore, the linguistic models, which are often at the heart of NLP tools, are also subject to change, not necessarily coordinated with the tools themselves. And once the first linguistic data with their metadata and annotations are available, the first analyses can begin to answer research questions, or the data can be explored with query interfaces.

3.2.1. Versioning of Corpora Using Git—The Case of Merlin

The Institute for Applied Linguistics (IAL) at Eurac Research is currently exploring how it can move towards a setup for more reproducibility in language learning research.

The first step that has been identified is to ensure that corpora are properly versioned. This makes it possible to refer to a specific version of a corpus that has been used as the basis for a published analysis even though the corpus has already been adapted and enhanced in the meantime. The first corpus that was moved into such a strictly versioned environment is the Merlin corpus [34]. The corpus is fully available on a publicly reachable on-premise GitLab installation (https://gitlab.inf.unibz.it/commul/merlin-platform/data-bundle, accessed on 22 April 2021). The repository is divided into several parts for the different formats in which the data are available and are accompanied by extensive documentation. The different versions of the corpus are realized as git tags, which are permanent references to specific points in the development history of the data. These tagged versions are also uploaded to the Eurac Research CLARIN Center (ERCC), which is the CLARIN-DSpace repository (see Section 3.1.2) hosted by the IAL, so they can be easily downloaded by less tech-savvy users. Another advantage is, of course, that this integration of the data into a CLARIN Center makes the metadata available to various search engines (e.g., the VLO or the OLAC search) and thus easy to find. All the data for a tagged version is available on both the ERCC and GitLab, each of these hosting platforms referencing the other. In both places, all versions are accompanied by a changelog explaining the changes between versions. On GitLab, the interested user can also make use of the integrated version diff to get more fine-grained information on the changes between versions. An important part of this setup is that all older versions of the corpus remain available to ensure reproducibility of earlier research. At the same time, all the versions are connected, and a user looking to use the corpus will prominently be made aware of which is the newest version.

3.2.2. Processing Pipelines and Reproducibility—KoKo

Another example is the Korpus Südtirol Project [50], an ongoing corpus linguistic initiative since 2005. One prominent subcorpus that has been used for several studies is the KoKo corpus (http://hdl.handle.net/20.500.12124/12, accessed on 22 April 2021) [11]. The ongoing nature of the project brings about the usual updates to the processing tools and utilities, for example, to the lemmatizer and part-of-speech tagger, the IMS TreeTagger [51], and to the corpus managing and querying tools, the IMS Open Corpus Workbench (CWB) [52]. However, it also brings about changes to interfaces or additions of tools. For example, we added the NoSketch Engine [53], a limited version of the software empowering the Sketch Engine service, and added ANNIS [5], a web browser-based search and visualization architecture for complex multilayer linguistic corpora with diverse types of annotation.

Any of these updates could bring changes to implicit default settings or to explicitly documented behavior, which, in turn, could cause changes in the processing and data representation, thus in the data used for subsequent analyses. To mediate such effects, it is necessary to track changes to all involved processing tools and utilities. Considering all of this from the perspective of comparability, it even becomes difficult to compare results from the same corpus but at different times, which, in turn, violates two pillars of research, namely repeatability and reproducibility. Naively, one could solve this problem by keeping different functional versions of the individual tools readily available. However, quickly, or at last, as soon as there exist numerous dependencies, this task becomes intractable.

Pipelines built on Docker images are a possible cure in this situation because they allow for encapsulated, fully functional units that can be kept for later re-use and ensure identical behavior. This is still a work in progress and not (yet) part of the infrastructure described in this paper. For some more detailed thoughts on this, see [54] and [55].

However, the KoKo corpus has one feature that requires special attention when implementing the paradigm for explicit versioning of corpora that has been described above: the corpus contains personal information for which the corpus creators asked the users for their consent to share the data, and this consent was explicitly requested for re-use in academic contexts. More generally, linguistic corpora often consist of personal data produced by individuals where both privacy and IPR concerns need to be considered. In addition, if not all the data can be made publicly available, there must be additional

access protection both on the side of the CLARIN-DSpace repository and on the side of GitLab. Although it is easy to have some data require a login with an academic account (for example, by using the CLARIN federated login) in CLARIN-DSpace, the GitLab repository should ideally not be made completely password-protected but have at least an openly available landing page that describes the corpus. At the IAL, this has been implemented for the KoKo corpus using git submodules where the main repository with the documentation and the overview of the various data formats is publicly accessible, and the actual data is in sub-repositories that require a login. Still, all license information and documentation are available without login. It is likely that more complex access scenarios will prove more difficult to map to a code hosting platform.

3.3. Ensuring Comparability and Reusability through FAIRness of the Integrated L1 and L2 Corpora

3.3.1. Findability

Within the Learner Corpus Infrastructure (LCI), Findability is being achieved by depositing all data in the institute's research data repository, Eurac Research CLARIN Center (ERCC) (https://clarin.eurac.edu, accessed on 22 April 2021). The repository uses the CLARIN-DSpace software (https://github.com/ufal/clarin-dspace, accessed on 22 April 2021), which has been developed by the Institute of Formal and Applied Linguistics, Charles University Prague within the LINDAT/CLARIAH-CZ project and has been adopted and refined by many other contributors. The software ensures that (meta)data are assigned a globally unique and persistent identifier (F1) and that the metadata clearly and explicitly include the identifier of the data they describe (F3). Regarding the need for rich metadata (F2), within the LCI, we are planning to develop a minimal set of metadata, but as this also touches upon the principles of Interoperability and Reusability, we will cover it in more detail there (see Sections 3.3.3 and 3.3.4). Moreover, because the ERCC is part of the European CLARIN infrastructure (https://www.clarin.eu, accessed on 22 April 2021) [56], every item that is being deposited will have its metadata automatically provided via OAI-PMH (https://www.openarchives.org/pmh/, accessed on 22 April 2021) in various formats, and this, in turn, is periodically being harvested by search engines such as the CLARIN Virtual Language Observatory (VLO) (https://vlo.clarin.eu, accessed on 22 April 2021) and the OLAC catalogue (http://search.language-archives.org/, accessed on 22 April 2021). These are two of the best-known search interfaces in the realm of (corpus) linguistics, and including our corpora there means that they can easily be found by interested researchers. Additionally, the Findability could be further increased by registering the corpora in lists of learner corpora. There is, for example, a corresponding CLARIN resource family for L2 corpora (https://www.clarin.eu/resource-families/L2-corpora, accessed on 22 April 2021).

3.3.2. Accessibility

Like Findability, a lot of the important requirements of the Accessibility principle are easily covered by depositing the data in a research data repository. By making the L1 and L2 learner corpora available through the ERCC, it is ensured that (meta)data are retrievable by their identifier using a standardized communications protocol (A1), and this protocol is open, free, and universally implementable (A1.1). Here the protocol is simply http(s). Moreover, the protocol allows for an authentication and authorization procedure where necessary (A1.2). As most of the corpora are only available for academic research, users will have to log in to get the data. CLARIN-DSpace provides easy authentication and authorization using the CLARIN federated identity, which means that users do not need to create a new account, but can simply log in using their university account, which also automatically shows that they are academic users. Regarding the principle that metadata is accessible, even when the data is no longer available, this is something that cannot be ensured via technological means. However, as data that has been deposited in the ERCC will get issued a persistent identifier and we believe that the "persistent" part should be

honored, there are regulations in place that even if data has to be removed for whatever reason, the metadata will stay online, including a note on when and why the data was removed (see also the corresponding point in the ERCC FAQ: https://clarin.eurac.edu/repository/xmlui/page/faq#how-to-delete, accessed on 22 April 2021).

3.3.3. Interoperability

To achieve Interoperability, the FAIR guiding principles for data stewardship suggest using a formal, accessible, shared, and broadly applicable language for knowledge representation for data and metadata (I1), to use vocabularies for data and metadata that themselves follow the FAIR principles (I2) and to include qualified references to other (meta)data (I3). The corpora collected and hosted by the IAL all use some formal and structured language for knowledge representation for the data. However, the original knowledge representation format that has been used usually was for all corpora some variation of a custom-tailored XML schema, with differing annotation schemes and error taxonomies and even different naming schemes for the same annotations using German, Italian or English denominators, depending on the focus of the project. Even though those data files and formats can thus be called formal and probably also accessible, thanks to the structured representation in XML, the schemas and vocabularies used are not well-documented, have only seen singular use and are not per se compatible with tools and applications for L1 and L2 learner corpus research—hence not broadly applicable. Moreover, metadata was originally not saved in structured, machine-actionable formats but in spreadsheets, XML headers or tab-separated text files. The used vocabularies were scarcely documented and usually had just a single project lifetime. To transform this data and metadata into FAIR compliant interoperable research data, we decided to perform the following steps. The data will be published in the ERCC (see above) offering data bundles with various formats for knowledge representation, including the originally used format (for documentation and replication activities) as well as additional formats obtained by conversion methods. The data will be provided in at least two different plain text versions, representing the originally composed text of the student/learner as well as a form-corrected version that allows automatic processing of the data to be more efficient. In terms of structured data formats, the provided data bundles will contain the existing custom-tailored XML version as well as a version in ANNIS format, which can be imported into the ANNIS corpus query software [5]. By offering this format, users can make use of the Salt and Pepper conversion framework [57] to convert the data into many other formats used for corpus linguistics in general and L1 and L2 learner corpus research in particular. The metadata for the whole corpus will be provided using the component metadata infrastructure (CMDI) format [58], a machine-actionable metadata format developed within the CLARIN community and supported by the CLARIN-DSpace software used in the ERCC repository. Additional document-level metadata (e.g., regarding the writer or the particular writing task) will be provided within the data bundles in the form of CMDI or tab-separated files. The provision of data in a TEI compliant format, which is increasingly used in learner corpus research (cf. [20]), is considered for future times, especially because it allows the inclusion of metadata on document level within the TEI header in the data files. Furthermore, the vocabulary used will be unified as much as possible over the five corpora provided via this infrastructure. For annotations, this will be supported by the definition of standard sets of minimal annotations that have proven useful over all five corpus projects. For metadata, this will be done by transforming existing metadata into a unified format using a standardized set of core metadata fields for learner corpora that is based on the suggestions made by Granger and Paquot [46,59]. Further research will deal with the challenge of how to make these vocabularies findable, accessible, interoperable and reusable as well. This is currently discussed within the CLARIN community in a specialized task force that one of the authors of this paper is a member of. Finally, to include qualified references to other metadata and data, all data are described on one central platform (https://www.porta.eurac.edu/, accessed on 22 April 2021) that links to

both data entries at research data repositories as well as to a corpus search interface. Both resources contain links to all available corpora with cross-references between themselves and to other (earlier) versions of the data or to related sub-corpora.

3.3.4. Reusability

The final and hardest step, however, is to provide reusable data. Although reusable data must be findable, accessible and interoperable as a prerequisite, further aspects are listed in the FAIR guidelines that ensure the final Reusability of the data. These recommendations concern, in particular, the description of data and metadata, using a "plurality of accurate and relevant attributes" (R1). This means the (meta)data should be released with a clear and accessible data usage license (R1.1) and associated with detailed provenance (R1.2). Furthermore, metadata and data should "meet domain-relevant community standards" (R1.3) [28]. Although corpus descriptions of the IAL corpora were previously published within single research papers or occasionally on project web pages, detailed knowledge about the individual resources was mostly limited to one or two people who were part of the project and/or the corpus creation team. The infrastructure should externalize this knowledge in a standardized and structured form, using the same corpus description templates for all resources. The templates are based on domain-dependent core metadata sets that build on suggestions having been made in the learner corpus community [59]. Although in theory, it would be best to provide as much metadata as possible, we aim to fulfil at least a minimal set of metadata on corpus and document level, including administrative metadata, corpus design metadata, corpus annotation metadata, text metadata, and learner metadata (categories as defined in [46]). The administrative metadata comprise general information on the data and its provenance, including information about the data collectors and all persons involved in the production and processing of the corpus. The corpus design and corpus annotation metadata describes the type and character of the data and gives information about the conducted annotation and processing activities. Text metadata contain information about the context of the text production and processing; among others, they specify the writing task, state when the data were collected/produced and in which way the collection process was set up. Finally, in the learner metadata, the writers, their background and their characteristics (e.g., mother tongue, proficiency in other languages) are described. To clear out eventual doubts about the usability of the published data that was previously only mentioned as "available for research purposes" within research or corpus description articles, we chose and assigned a unified license for all five corpora (EULA-CLARIN-ACA-BY-NC-NORED). The license can be found in the respective CLARIN-DSpace collection for each corpus and on GitLab (e.g., https://gitlab.inf.unibz.it/commul/koko/data/bundle/-/blob/master/EULA-CLARIN-ACA-BY-NC-NORED.md, accessed on 22 April 2021). The possibility to redistribute the data using this license was provided by former actions ensuring legal and ethical use of the data, that is, the acquisition of an explicit user consent of the writers or their parents in case they were still minors. The license text is identical for each corpus except for the corpus name.

4. Discussion

As already stated, many L1 and L2 corpora are currently collected in—especially when it comes to non-English-based research—rather small projects, which regardless of their size require considerable amounts of time and effort for data collection and corpus creation activities. However, making these research data available in a FAIR way, with standardized and reasoned methods, would contribute substantially to the advancement of the field and would answer current demands in transparency, reproducibility and reusability following the standards set for open science (see Section 1).

Even though there is an increasing awareness of the need to open up L1 and L2 learner corpora and tools for the wider research community [41], little effort has been put into ensuring that new resources will be created and distributed per default in a FAIR way.

This is partly because most L1 and l2 corpus projects are research projects focused on their own research questions. Their priority is to ensure that data collection is as intuitive to the researcher as possible and that the data they produce fits the aims of the investigation. Only afterwards—if even—thoughts are put into what happens to the data after the project life span.

In the work presented in this article, we have attempted to make several corpora available to a wider audience after the original projects in which these corpora were collected had been completed. We integrated them in a newly built research infrastructure based on solutions from the more general research infrastructure community and tailored for L1 and L2 corpus resources and put efforts into ensuring that the corpora comply as much as possible with the FAIR guiding principles for data stewardship as well as with our own considerations regarding reproducibility concerns.

While referring to established workflows and solutions for the general research infrastructure community, we created an L1 and L2 corpus infrastructure that solves the main issues for making the data findable and accessible (see Section 2.2). However, when it comes to interoperability and reusability, domain-specific solutions need to be defined. In the case of L1 and L2 corpora, we tried to address this by finding solutions that are appropriate for several different L1 and L2 corpora that have been collected throughout various research projects.

Our attempts on making those corpora available to a wider audience in retrospect have, however, shown that this activity is very time-consuming and tedious and often underlies constraints that cannot be overcome, since data formats, metadata sets and usage consents that hugely define what is possible and what not, had already been set before. Although newly created resources could solve a lot of the issues of discoverability and availability for the wider research community and partly even standardization by forward-looking research data management, the publication of data from past projects can sometimes be impossible due to legal and ethical issues, for example, in the case that no consent for (wide) publication was obtained from the subjects and most of them are impossible to re-contact because data has been anonymized, or due to undocumented data formats that cannot be converted to interoperable formats. This has been discussed before (e.g., ref. [43]) and can only be underlined by our results.

However, for new resources to be designed and created with subsequent data sharing in mind, more than a higher awareness for the issues related to it and a stronger commitment to making data available are needed. The lack of coherence between past research projects that take part in L1 and L2 learner corpus creation that we identified in Section 1 is currently sustained by a lack of best-practice templates and model solutions for data management in L1 and L2 learner corpora that have currency for various projects (see Section 2).

The solutions we found for integrating the resources of the IAL in an L1 and L2 learner corpus infrastructure (see Section 3) could be a preliminary contribution to filling this gap. However, further work is needed, involving more stakeholders from L1 and L2 corpus creation and extending metadata schemes and proposed data formats used here to also fit the requirements of other types of L1 and L2 corpora that have not been considered so far (e.g., spoken or multi-modal corpora, corpora collected within official language testing frameworks, or from translators, etc.).

5. Conclusions

The research community as a whole widely agrees that data produced during scientific research is a very valuable resource, and making it available following the FAIR principles should be seen as the ideal towards which all researchers should strive within their projects. This has the immediate consequence that the data is available for colleagues to reproduce research results—one of the most important scientific principles—and especially in the domain of (corpus) linguistics, a laboriously collected and well-curated dataset can be very valuable as a resource for further research, and well-documented methods and open tools

can help further research by being able to be readily applied to novel data thus making comparisons possible.

In recent years, substantial efforts have been made in Europe as well as worldwide to construct robust digital infrastructures for research data that follow the FAIR principles. As these infrastructures are aimed at serving the scientific community as a whole, a lot of the tools and methods that were developed and made available are domain agnostic to a certain degree and focus on the FAIR principles of Findability and Accessibility, which can be achieved more or less in the same way whether it pertains to linguistics, astronomy or genetics.

The research infrastructures community can only raise awareness for the importance of clear and transparent research data management and FAIR data sharing while providing the technical means for long-term preservation, aggregation, and distribution on a large scale. The solutions designed and provided by general research infrastructure players like, for example, CLARIN or META-SHARE [60] typically must span various research communities, thereby being relatively flexible and permissive in terms of type and specifics of the use cases covered. Thus, the generic solutions provided can help to guarantee that the first two principles of FAIR are ensured since they can be solved in mostly domain-agnostic ways. For data to be fully FAIR compliant community-specific solutions need to be found, discussed and agreed upon.

Although there were some domain-specific efforts in creating research infrastructures for L1 and L2 learner corpora in the past (e.g., ref. [21,61,62]), the infrastructures created so far usually focus on a single or a small family of L1 or L2 corpora, making available data and tools that were developed during a project. A wider adaption of methods and formats by other projects has rarely happened so far [20,44]. Requests for more standardization of L1 and L2 learner corpora as in [20,44] show that the community sees the problem of highly idiosyncratic corpora and methodology and acknowledges the need for working together to create common standards to facilitate interoperability but so far, no forum has emerged that could push towards this direction.

The Learner Corpus Infrastructure (LCI) presented in this paper has a two-fold aim. First, it aims to create a common standard within the Institute for Applied Linguistics at Eurac Research by taking all the corpora collected there within the last 15 years and making them undergo a process of FAIRification. In the end, all corpora will be available in a small number of standardized file formats, and the texts will be annotated using a consistent annotation scheme and documented using a unified metadata scheme that was evaluated on various resources and went through many iterations. Furthermore, the data is made available using the specialized research data repository software CLARIN-DSpace. In this, and by making the metadata available in CMDI format, the whole LCI is embedded in the European CLARIN community, ensuring wide discoverability and easy access through the CLARIN authentication and authorization infrastructure. Finally, the whole data is presented through a user-friendly portal which provides easy access to all the corpora, including the possibility to search them using a standardized search interface.

The other aim of this effort is to use the LCI as a testbed for possible steps to make the whole field of learner corpus research more FAIR. We show how infrastructure solutions developed for more general purposes can contribute by building a base for domain-specific infrastructures and discuss what further domain-specific steps are needed for comparable, reusable and reproducible data. The corpora integrated cover a relatively wide variety of cases with both L1 and L2 learner corpora and monolingual as well as multilingual ones. We hope that the infrastructure itself and articles such as this where we explain our reasoning behind the various steps and highlight the problematic issues that still remain can help the community to further the discussion on standardization of learner corpora and how the field as a whole can become more FAIR.

As we have outlined, there are several issues that need community involvement to advance reproducibility and reusability in the field of L1 and L2 learner corpora. We have also shown that there is widespread interest in creating research infrastructures both within the

domain as well as spanning various domains as a general trend in research. This work can help with some of the requirements for reusable, comparable and reproducible data and research results. However, other aspects should be further pursued involving a wide variety of key actors in the community. To achieve this, networking activities need to be initiated that could be realized on a European level, for example, through the creation of a COST Action (https://www.cost.eu/cost-actions/what-are-cost-actions/, accessed on 22 April 2021) or an Erasmus+ Knowledge Alliance (https://ec.europa.eu/programmes/erasmus-plus/programme-guide/part-b/three-key-actions/key-action-2/knowledge-alliances_en, accessed on 22 April 2021), both of which have their focus on strengthening existing research communities by providing funds for networking and knowledge exchange.

Furthermore, one of the strategic priorities of CLARIN ERIC is to promote standardization and FAIR data within more linguistic domains and to this effectively, they provide funding, for example, for organizing workshops with these topics (https://www.clarin.eu/content/clarin-workshops, accessed on 22 April 2021). An event of that kind, the CLARIN workshop on "Interoperability of Second Language Resources and Tools" has taken place in Gothenburg in 2017 [19], but to reach the goals outlined here, a steadier exchange is needed. A further possibility within the CLARIN community would be to set up a distributed CLARIN Knowledge Center (https://www.clarin.eu/content/knowledge-centres, accessed on 22 April 2021) for L1 and L2 learner corpora. This could serve as an anchor for future work in this regard and help the community to gain higher visibility, even beyond its own domain.

Author Contributions: Conceptualization, A.K., J.-C.F. and E.W.S.; methodology, A.K., J.-C.F. and E.W.S.; software, A.K. and E.W.S.; validation, A.K., J.-C.F. and E.W.S.; formal analysis, A.K., J.-C.F. and E.W.S.; investigation, A.K., J.-C.F. and E.W.S.; resources, A.K., J.-C.F. and E.W.S.; data curation, A.K., J.-C.F. and E.W.S.; writing—original draft preparation, A.K., J.-C.F. and E.W.S.; writing—review and editing, A.K., J.-C.F. and E.W.S.; visualization, A.K., J.-C.F. and E.W.S.; supervision, A.K., J.-C.F. and E.W.S.; project administration, A.K., J.-C.F. and E.W.S.; funding acquisition, A.K., J.-C.F. and E.W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is or soon will be available in a publicly accessible repository. See Section 2.1 for details.

Acknowledgments: We thank Darja Fišer for her help with the article submitted to the ICTeSSH2020 conference proceedings on which this article is based and for valuable comments on this article itself. We also thank CLARIN ERIC for financing this submission.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Granger, S. Learner corpora in foreign language education. In *Language and Technology. Encyclopedia of Language and Education*; Thorne, S., May, S., Eds.; Springer: Cham, Switzerland, 2017; pp. 427–440._33. [CrossRef]
2. Granger, S.; Hung, J.; Petch-Tyson, S. *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching*; John Benjamins Publishing: Amsterdam, The Netherlands, 2002; Volume 6.
3. Schmidt, T. EXMARaLDA and the FOLK tools—Two toolsets for transcribing and annotating spoken language. In *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC'12)*; Declerck, T., Choukri, K., Calzolari, N., Eds.; European Language Resources Association (ELRA): Miyazaki, Japan, 2012; pp. 236–240.
4. Kilgarriff, A.; Marcowitz, F.; Smith, S.; Thomas, J. Corpora and Language Learning with the Sketch Engine and SKELL. *Rev. Fr. Linguist. Appl.* **2015**, *XX*, 61–80. [CrossRef]
5. Krause, T.; Zeldes, A. ANNIS3: A new architecture for generic corpus query and visualization. In *Digital Scholarship in the Humanities*; The Oxford University Press: Oxford, UK, 2016; Volume 31, pp. 118–139. [CrossRef]
6. Janssen, M. TEITOK: Text-faithful annotated corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*; Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., et al., Eds.; European Language Resources Association (ELRA): Miyazaki, Japan, 2016; pp. 4037–4043.

7. Okinina, N.; Nicolas, L.; Lyding, V. Transc&Anno: A graphical tool for the transcription and on-the-fly annotation of handwritten documents. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*; Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., et al., Eds.; European Language Resources Association (ELRA): Miyazaki, Japan, 2018.
8. Volodina, E. Korp Searches in Second Language Data—Språkbanksbloggen. Available online: https://spraakbanken.gu.se/blogg/index.php/2020/06/17/korp-searches-in-second-language-data/ (accessed on 22 April 2021).
9. Centre for English Corpus Linguistics. Learner Corpora around the World. 2020. Available online: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (accessed on 22 April 2021).
10. Fišer, D.; Lenardič, J.; Erjavec, T. CLARIN's key resource families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*; Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., et al., Eds.; European Language Resources Association (ELRA): Miyazaki, Japan, 2018.
11. Abel, A.; Glaznieks, A.; Nicolas, L.; Stemle, E.W. KoKo: An L1 learner corpus for german. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014; pp. 2414–2421.
12. Nesselhauf, N. Learner corpora and their potential for language teaching. In *How to Use Corpora in Language Teaching*; John Benjamins Publishing: Amsterdam, The Netherlands, 2004; Volume 12, pp. 125–156.
13. Gilquin, G.; Granger, S. From design to collection of learner corpora. In *The Cambridge Handbook of Learner Corpus Research*; Cambridge University Press: Cambridge, UK, 2015; Volume 3, pp. 9–34.
14. Hunston, S. Corpus compilation collection strategies and design decisions. In *Corpus Linguistics: An International Handbook*; De Gruyter Mouton: Berlin, Germany, 2009; Volume 2, pp. 154–168.
15. Gries, S.T.; Newman, J. Creating and using corpora. *Res. Methods Linguist.* **2014**, *2*, 257–287. [CrossRef]
16. Lenardič, J.; Tiedemann, T.L.; Fišer, D. *Overview of L2 Corpora and Re-Sources 2.0*; Technical Report; CLARIN: Utrecht, The Netherlands, 2018.
17. Gries, S. On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *J. Second Lang. Stud.* **2018**, *1*, 276–308. [CrossRef]
18. Paquot, M.; Plonsky, L. Quantitative research methods and study quality in learner corpus research. *Int. J. Learn. Corpus Res.* **2017**, *3*, 61–94. [CrossRef]
19. Volodina, E.; Tenfjord, K.; Mikelic Preradovic, N.; Janssen, M.; Lindström Tiedemann, T.; Ragnhildstveit, S. Workshop on Interoperability of L2 Resources and Tools | sweclarin.se. Available online: https://sweclarin.se/swe/workshop-interoperability-l2-resources-and-tools,2017 (accessed on 22 April 2021)
20. Stemle, E.W.; Boyd, A.; Janssen, M.; Lindström Tiedemann, T.; Mikelić Preradović, N.; Rosen, A.; Rosén, D.; Volodina, E. Working together towards an ideal infrastructure for language learner corpora. In Proceedings of the Widening the Scope of Learner Corpus Research Selected Papers from the Fourth Learner Corpus Research Conference 2017, Bolzano/Bozen, Italy, 5–7 October 2017; pp. 437–478.
21. Volodina, E.; Megyesi, B.; Wirén, M.; Granstedt, L.; Prentice, J.; Reichenberg, M.; Sundberg, G. A friend in need?: Research agenda for electronic Second Language infrastructure. In Proceedings of the Sixth Swedish Language Technology Conference (SLTC), Umeå, Sweden, 17–18 November 2016.
22. Glaznieks, A.; Abel, A.; Lyding, V.; Nicolas, L.; Stemle, E.W. Establishing a Standardised Procedure for Building Learner Corpora. *Apples J. Appl. Lang. Stud.* **2014**, *8*, 5–20.
23. Treloar, A. The Research Data Alliance: Globally co-ordinated action against barriers to data publishing and sharing. *Learn. Publ.* **2014**, *27*, 9–13. [CrossRef]
24. Moskovko, M. Intensified role of the European Union? European Research Infrastructure Consortium as a legal framework for contemporary multinational research collaboration. In *Big Science and Research Infrastructures in Europe*; Edward Elgar Publishing: Surrey, UK, 2020.
25. Ayris, P.; Berthou, J.Y.; Bruce, R.; Lindstaedt, S.; Monreale, A.; Mons, B.; Murayama, Y.; Södergård, C.; Tochtermann, K.; Wilkinson, R. Realising the European Open Science Cloud. First Report and Recommendations of the Commission High Level Expert Group on the European Open Science Cloud. 2016. Available online: file:///C:/Users/MDPI/AppData/Local/Temp/RealisingtheOpenScienceCloud-2.pdf (accessed on 22 April 2021).
26. Veršić, I.I.; Ausserhofer, J. Social sciences, humanities and their interoperability with the European Open Science Cloud: What is SSHOC? *Mitt. Ver. Österreichischer Bibl. Bibl.* **2019**, *72*, 383–391.
27. European Language Resources Association (ELRA). *Social Sciences and Humanities Pathway Towards the European Open Science Cloud*; European Language Resources Association (ELRA): Paris, France, 2020; [CrossRef]
28. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et.al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 1–9. [CrossRef]
29. Barbot, L.; Biller, T.; Broeder, D.; Dekker, R.; Durco, M.; Vipavc, I.; Willems, M. Agile development of the SSH open marketplace: User workshop. In *ITM Web of Conferences*; EDP Sciences: Ulis, Paris, 2020; Volume 33, p. 04001.
30. de Jong, F.M.G.; Maegaard, B.; De Smedt, K.; Fišer, D.; Van Uytvanck, D. CLARIN: Towards FAIR and responsible data science using language resources. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.

31. Abel, A.; Vettori, C.; Wisniewski, K. *KOLIPSI: Gli Studenti Altoatesini e la Seconda Lingua; Indagine Linguistica e Psicosociale= KOLIPSI: die Südtiroler SchülerInnen und die Zweitsprache; Eine Linguistische und Sozialpsychologische Untersuchung*; Eurac Research: Bolzano/Bozen, 2012. Available online: https://www.researchgate.net/publication/259453091_Gli_studenti_altoatesini_e_la_seconda_lingua_indagine_linguistica_e_psicosociale_Die_Sudtiroler_SchulerInnen_und_die_Zweitsprache_eine_linguistische_und_sozialpsychologische_Untersuchung_Volume_1_-_Ba (accessed on 22 April 2021).
32. Vettori, C.; Abel, A. (Eds.) KOLIPSI II Gli studenti altoatesini e la seconda lingua: indagine linguistica e psicosociale. In *Die Südtiroler SchülerInnen und die Zweitsprache: Eine Linguistische und Sozialpsychologische Untersuchung*; Eurac Research: Bolzano/Bozen, 2017. Available online: https://bia.unibz.it/discovery/delivery?vid=39UBZ_INST:ResearchRepository&repId=12235320180001241#13235268510001241 (accessed on 22 April 2021).
33. Abel, A.; Glaznieks, A.; Nicolas, L.; Stemle, E. An extended version of the KoKo German L1 Learner corpus. In Proceedings of the Third Italian Conference on Computational Linguistics, Napoli, Italy, 5–6 December 2016.
34. Boyd, A.; Hana, J.; Nicolas, L.; Meurers, D.; Wisniewski, K.; Abel, A.; Schöne, K.; Štindlová, B.; Vettori, C. The MERLIN corpus: Learner language and the CEFR. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 14), Reykjavik, Iceland, 26–31 May 2014; pp. 1281–1288.
35. Zanasi, L.; Stopfner, M. Rilevare, osservare, consultare. Metodi e strumenti per l'analisi del plurilinguismo nella scuola secondaria di primo grado. *La didattica delle lingue nel nuovo millennio* **2018**, 135–148. Available online: https://edizionicafoscari.unive.it/media/pdf/books/978-88-6969-228-4/978-88-6969-228-4-ch-01_ALK6Jr7.pdf (accessed on 22 April 2021).
36. Granger, S.; Dagneaux, E.; Meunier, F.; Paquot, M. *International Corpus of Learner English*; Presses Universitaires de Louvain: Louvain-la-Neuve, Belgium, 2009.
37. Tenfjord, K.; Meurer, P.; Hofland, K. The ASK corpus-a language learner corpus of norwegian as a second language. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, 22–28 May 2006; pp. 1821–1824.
38. Rosen, A.; Hana, J.; Vidová Hladká, B.; Jelínek, T.; Škodová, S.; Štindlová, B. *Compiling and Annotating a Learner Corpus for a Morphologically Rich Language: {CzeSL}, a Corpus of Non-Native {Czech}*; Nakladatelství Karolinum, 2020. Available online: http://hdl.handle.net/20.500.11956/123103 (accessed on 22 April 2021).
39. Blanchard, D.; Tetreault, J.; Higgins, D.; Cahill, A.; Chodorow, M. TOEFL11: A corpus of non-native English. *ETS Res. Rep. Ser.* **2013**, *2*, 15. [CrossRef]
40. Mons, B.; Neylon, C.; Velterop, J.; Dumontier, M.; da Silva Santos, L.O.B.; Wilkinson, M.D. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Inf. Serv. & Use* **2017**, *37*, 49–56. [CrossRef]
41. Lindström, T.; Lenardič, J.; Fišer, D. L2 learner corpus survey–Towards improved verifiability, reproducibility and inspiration in learner corpus research. In Proceedings of the CLARIN Annual Conference 2018, Pisa, Italy, 8–10 October 2018; pp. 146–150.
42. Van Uytvanck, D.; Stehouwer, H.; Lampen, L. Semantic metadata mapping in practice: the Virtual Language Observatory. In *LREC 2012: 8th International Conference on Language Resources and Evaluation*; European Language Resources Association (ELRA): Paris, France, 2012; pp. 1029–1034.
43. Megyesi, B.; Granstedt, L.; Johansson, S.; Prentice, J.; Rosén, D.; Schenström, C.J.; Sundberg, G.; Wirén, M.; Volodina, E. Learner corpus anonymization in the age of gdpr: Insights from the creation of a learner corpus of swedish. In Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL 2018) at SLTC, Stockholm, Sweden, 7 November 2018; Linköping University Electronic Press: Linköping, Sweden, 2018; pp. 47–56.
44. Volodina, E.; Janssen, M.; Tiedemann, T.L.; Preradović, N.M.; Ragnhildstveit, S.; Tenfjord, K.; de Smedt, K. Interoperability of Second Language Resources and Tools. In Proceedings of the CLARIN Annual Conference 2018, Pisa, Italy, 8–10 October 2018; pp. 90–94.
45. Chiarcos, C.; Nordhoff, S.; Hellmann, S. *Linked Data in Linguistics*; Springer: New York, NY, USA, 2012.
46. Granger, S.; Paquot, M. Towards standardization of metadata for L2 corpora. In Proceedings of the workshop on Interoperability of Second Language Resources and Tools, Gothenburg, Sweden, 6–8 December 2017.
47. Wittenburg, P.; Van Uytvanck, D.; Zastrow, T.; Strak, P.; Broeder, D.; Schiel, F.; Boehlke, V.; Reichel, U.; Offersgaard, L. *CLARIN B Centre Checklist*; Technical Report CE-2013-0095; Clarin Eric: Utrecht, The Netherlands, 2018.
48. Eskevich, M.; de Jong, F.; König, A.; Fišer, D.; Van Uytvanck, D.; Aalto, T.; Borin, L.; Gerassimenko, O.; Hajic, J.; van den Heuvel, H.; et al. CLARIN: Distributed language resources and technology in a European infrastructure. In *Proceedings of the 1st International Workshop on Language Technology Platforms*; European Language Resources Association (ELRA): Marseille, France, 2020; pp. 28–34.
49. Nicolas, L.; Stemle, E.; Glaznieks, A.; Abel, A. A Generic Data Workflow for Building Annotated Text Corpora. *Stud. Learn. Corpus Linguist. Res. Appl. Foreign Lang. Teach. Assess.* **2015**, *190*, 337–351. [CrossRef]
50. Abel, A.; Anstein, S. Korpus südtirol—Varietätenlinguistische untersuchungen. In *Korpora in Lehre und Forschung*; Abel, A., Zanin, R., Eds.; Bozen-Bolzano University Press: Bozen, Italy, 2011; pp. 29–54.
51. Schmid, H. Improvements in part-of-speech tagging with an application to German. In Proceedings of the ACL SIGDAT-Workshop, Dublin, Ireland, 30 June 1995; pp. 47–50.
52. Evert, S.; Hardie, A. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In Proceedings of the Corpus Linguistics 2011, Birmingham, UK, 20–22 July 2011.
53. Rychlý, P. Manatee/Bonito—A modular corpus manager. In *First Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2007)*; Masaryk University: Brno, Czech Republic, 2007; pp. 65–70.

54. König, A.; Stemle, E.W.; Moreira, A.; Elbers, W. Technical solutions for reproducible research. In *Selected Papers from the CLARIN Annual Conference 2019*; Simov, K., Eskevich, M., Eds.; Linköping University Electronic Press: Linköping, Sweden, 2020; Volume 172, pp. 66–74. [CrossRef]
55. Branco, A.; Calzolari, N.; Vossen, P.; Van Noord, G.; Van Uytvanck, D.; Silva, J.; Gomes, L.; Moreira, A.; Elbers, W. 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*; European Language Resources Association: Paris, France, 2020; pp. 5539–5545.
56. Krauwer, S.; Hinrichs, E. The CLARIN research infrastructure: resources and tools for e-humanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*; European Language Resources Association: Paris, France, 2014; pp. 1525–1531.
57. Druskat, S.; Gast, V.; Krause, T.; Zipser, F. Corpus-tools. org: An interoperable generic software tool set for multi-layer linguistic corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*; European Language Resources Association: Paris, France, 2016; pp. 4492–4499.
58. Broeder, D.; Windhouwer, M.; Van Uytvanck, D.; Goosen, T.; Trippel, T. CMDI: A component metadata infrastructure. In Proceedings of the Workshop on Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR, Istanbul, Turkey, 22 May 2012.
59. Granger, S.; Paquot, M. Core metadata for learner corpora: Eraft 1.0. In Proceedings of the workshop on Interoperability of Second Language Resources and Tools, Gothenburg, Sweden, 6–8 December 2017.
60. Piperidis, S. The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 21–27 May 2012; pp. 36–42.
61. Alfter, D.; Borin, L.; Pilán, I.; Tiedemann, T.L.; Volodina, E. Lärka: From language learning platform to infrastructure for research on language learning. In Proceedings of the CLARIN Annual Conference 2018, Pisa, Italy, 8–10 October 2018; pp. 53–56.
62. Darģis, R.; Auziņa, I.; Levāne-Petrova, K.; Kaija, I. Quality focused approach to a learner corpus development. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 392–396.