Creating a learner corpus infrastructure: Experiences from making learner corpora available

Jennifer-Carmen Frey[1,*], *Alexander* König[2], and *Darja* Fišer[34]

[1]Eurac Research, Institute for Applied Linguistics, 39100 Bolzano, Italy
[2]CLARIN ERIC, 3512 BS Utrecht, The Netherlands
[3]University of Ljubljana, Department of Translation, 1000 Ljubljana, Slovenia
[4]Department of Knowledge Technologies, Jozef Stefan Institute, 1000 Ljubljana, Slovenia

Abstract. With language resources being collected in many - also small - projects in learner corpus research with considerate amounts of time and effort spent in this activity, making these types of data available in a FAIR way, with standardized and reasoned methods, would contribute substantially to the advancement of the field. Additionally, it would answer current demands in transparency, replicability and reusability. In this article, we discuss some of the challenges when making learner corpora FAIR and report from experiences in fulfilling this aim while creating a learner corpus infrastructure at a research institution hosting five different learner corpora.

1 Introduction

Learner corpus research uses "electronic collections of natural or near natural foreign or second language learner texts" [1] to investigate the dynamics and outcomes of language learning processes on empirical data (cf. [2]). As such, this specialised domain that has grown into a well-established research field with a solid community[1] was not able to exclusively use general-purpose digital infrastructures already provided by third parties, but has engaged in the creation of their own, custom-tailored solutions, including the collection of their data and the creation of the aforementioned learner corpora as language resources for their endeavours. In its three decades of existence, a remarkable number of learner corpora has been collected and referred to in the studies presented in this field (see for example the ones listed by the Centre for English Corpus Linguistics at the University of Louvain [3]).

In addition, there are also corpora illustrating native language speakers' competences in their first language (L1) as opposed to texts in the writers' second or foreign language (L2). Such corpora, elicited for example in educational settings, are used to research literacy development and academic writing. Given that both types of texts are produced by a (not yet) fully proficient writer, and whose language, writing and text competence is still in development, both categories are sometimes referred to as learner corpora in a broader sense (e.g. in [4, 5, 6]). Although research questions and interested disciplines are slightly different, both

* Corresponding author: jennifer.frey@eurac.edu
[1] See also the large number of members of the Learner Corpus Association that was founded to provide an international and interdisciplinary forum for learner corpus research (https://www.learnercorpusassociation.org/)

data types have similar characteristics, including language that deviates from the standard variety, and consequently have a need for normalisation and/or error annotation; as well as the need for strict privacy policies and explicit user consents[2]. Hence, L2 learner corpora and L1 corpora used to analyse literacy development have similar technical requirements for data collection, processing and curation. Both of them are usually laboriously collected through interaction with language learning institutions, often elicited through pen and paper tests and subsequent transcription and are "systematic" [9] and "assembled according to explicit design criteria" [1], which means that their composition, creation and representation are project-dependent and tailored to the envisioned purpose of the resource [10]. Moreover, before being used, the corpora usually undergo various manual or semi-manual annotation processes in which additional data is added to the plain text material, providing researchers with (searchable) frequency information on specific linguistic phenomena that are of their interest.

The variety of individual solutions for the aforementioned steps leads to the fact that the learner corpora that have been created in individual research entities are all very diverse, not only in their content but also in their design and implementation, and while many of them are only available for the members of a particular research entity or even proprietary to a single researcher, even for the ones that are available to the general public, the interoperability and comparability of different language resources is rather low [11]. The interoperability of learner corpora is hampered by a wide range of formats and languages for knowledge representation, vastly diverging vocabularies for metadata and taxonomies for linguistic annotation as well as processing and analysis tools [12]. As a consequence, this greatly inhibits the comparison of results, their reproducibility or replicability, as well as the reusability or even hypothesized aggregation of the data [13, 14]. While transparency, reproducibility and replicability of results would answer recent demands for more rigor in learner corpus research [15, 16], making laboriously collected data reusable would allow the data collector as well as the greater academic public to exploit the resources more thoroughly, saving resources and time for further projects and advancing the field significantly (see also [14]). This has set the grounds for the recent attention on standardization in learner corpus research (e.g. [17], see also [13] who identified a strong need for interaction between research institutions involved in building learner corpora, proposing the creation of a European network dedicated to this issue) and the interest in the creation of digital infrastructures that make learner corpora and corresponding tools available to the academic public (e.g. [18, 7]). However, making learner corpus data findable, available and reusable in practice is subject to many considerations and an appropriate infrastructure is needed to approach this goal.

In this article we make use of the FAIR approach to data management by Wilkinson et al. [19] that is based on the principles of Findability, Accessibility, Interoperability and Reusability of research data as a guide towards implementing a digital infrastructure for making learner corpus data available to the wider research community. Although previous work on data management and interoperability was predominantly directed to L2 learner corpora, we believe that most of the problems discussed apply to L1 corpora for literacy development as well and will therefore include this type of learner corpus in a broader sense in our considerations. The article will first introduce some general issues with the FAIR aspects when applied to learner corpora and then present our experiences when trying to apply the FAIR principles in building a learner corpus research infrastructure at a European research institute that has gathered a significant amount of learner corpus data over the last 15 years.

2 Challenges when making learner corpus data available

[2] For a detailed discussion on similarities of L1 and L2 corpora and arguments for calling both of them learner corpora see [7, 8].

With ever more learner corpora being compiled and an increasing awareness of the needs to open them up for the broader research community [14], authors of such corpora are facing the following key obstacles:

(1) the legal and ethical issues related to informed consent and privacy protection of the informants are substantial, especially in projects focusing on minors, minorities or speech impairments, and in projects compiling multimodal data (e.g. voice, face) and cannot be resolved in arrears, so curation and publication of corpora from past projects is in most cases not feasible;

(2) the lacking infrastructural support for ensuring corpora are interoperable [12], such as a common set of metadata for learner corpora, a common error annotation schema, and a common encoding format, hinder reuse of existing corpora and comparative research.

On the other end, the main bottlenecks for potential users of the existing learner corpora are the following:

(1) Information about the corpora that already exist is scattered over numerous project and interest group websites, publications and repositories, making it difficult for users to look them up efficiently and even more difficult to access them for querying or download;

(2) Documentation about corpus design, preprocessing, annotation (schema as well as accuracy), encoding is not readily available or user-friendly;

(3) Different corpora are made available in different corpus querying tools, which offer different functionalities and calculate corpus statistics differently, making it difficult to compare results across corpora.

This is why research infrastructure support spanning the full FAIR spectrum is of crucial importance to make learner corpora available and case reusable, which will be outlined in the remainder of this section.

2.1 Findability of learner corpora

In 2018, Lindström Tiedemann et al. [14] conducted a survey on existing L2 learner corpora, identifying the Findability of L2 learner corpora as one of the major issues for this type of language resources (next to insufficient metadata availability and documentation). 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)[3], a well-known search engine for language resources in linguistics that allows to search on various metadata fields [20]. On the one hand, the lacking Findability of the language resources is certainly due to the many corpora that are produced more or less privately 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 [14]). However, it was also found to be hampered by missing metadata and the lack of the technical prerequisites to search for learner corpus-specific categories in the search interface.

2.2 Accessibility of learner corpora

According to the evaluation of Lindström Tiedemann et al. [14], only 23 out of the 31 L2 learner corpora that were findable, were also available for download or querying and thus accessible in some way. In many cases, this can be traced back to the aforementioned legal constraints of the data including missing or too narrow user consents or missing anonymisation/pseudonymisation of the data, that most often has to be done manually (see e.g. [21]). As legal requirements for publishing the corpus to a broader audience are most often country-dependent and contradicting [23], this is a serious issue in learner corpus research that often

[3] https://vlo.clarin.eu/

deals with rather sensitive data of minors (see also [12]). However, according to the FAIR principles the metadata and data should be available via a standardised protocol that is open and allows for authentication, offering at least the metadata of the corpora, once the corpus data is no longer available. So, even though the data might not be available, due to legal and ethical constraints it would be crucial to provide at least the metadata for the corpus in a standardised accessible way, so that researchers are able to learn from previous corpus creation project, compare their research and designs and potentially draw from existing knowledge while making resources and studies more comparable (cf. also [14])

2.3 Interoperability of learner corpora

The Interoperability of learner corpora has received increasing attention in the last years as researchers have identified the need for more consensus and international collaboration to make learner corpus research more comparable (e.g. [12, 13]). Volodina et al. [13], 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."

When looking at the formats used for knowledge representation (i.e. structural interoperability [18]), no "formally established standards" have been defined for the domain of learner corpora. For the language data itself, many researchers use the Text Encoding Initiative (TEI)[4] XML format[5] (often in combination with the IMS corpus workbench query language[6]) or PAULA/XML[7] (in combination with ANNIS[8] for corpus query) (see [13]). 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 [13]. 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 (cf. [19]).

In terms of conceptual interoperability [22], Volodina et al. [13] 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 established yet, as this would need directed activities in making those accepted vocabularies also explicitly findable, accessible, interoperable and reusable, including assigning them persistent identifiers, making them available via standardised protocols (as compared to privately curated and distributed lists), presenting them in a formal language for knowledge representation and describing them thoroughly.

Finally, the FAIR guidelines recommend making qualified references to other relevant data and metadata, which would make comparison between similar or even related data easier. However, few learner corpora are described next to each other, with comparable infor-

[4] https://tei-c.org/
[5] Stemle et al. [8] therefore names it as "de-facto" standard.
[6] http://cwb.sourceforge.net/
[7] http://www.sfb632.uni-potsdam.de/en/paula.html
[8] https://corpus-tools.org/annis/

mation on the same platforms or can be queried with the same query interface. Usually corpora are described on their own, individual webpages or come with their own search interfaces that very rarely give direct links to other data.

2.4 Reusability of learner corpora

Without extensive documentation of data with relevant metadata, learner corpora cannot be reused. Without detailed information about the purpose and circumstances of data collection, the background of the writers (e.g. L1, age, proficiency, other L2s) and the writing task itself (target language, genre, writing prompts etc.), it is impossible to judge whether a corpus is a suitable resource for one's research questions [10]. 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" [13]. Lindström Tiedemann et al. [14] and Granger and Paquot [17] point out that crucial metadata information is frequently missing, proposing the creation of a standardized core metadata set that can be reused 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 [14] who only found licensing information for 28 out of 31 findable corpora).

3 An agenda for FAIR learner corpora at Eurac Research

Having studied learner language and literacy development in multilingual areas and for pluricentric languages through various research projects, the Institute for Applied Linguistics at Eurac Research (IAL) has collected a substantial amount of learner data, both in the form of traditional L2 learner productions as well as in the form of corpora illustrating the development of L1 competences of students. This data, although related to individual and concluded research projects, could receive ongoing attention and potentially inform new research, when made available to the research community in an appropriate form that allows the reuse of such laboriously collected research data. For this purpose, the creation of a Learner Corpus Infrastructure (LCI) was initiated, which should provide a fruitful environment for future learner corpus research at the institute and allow the continued exploitation, maintenance, dissemination and preservation of the data including access for internal and external researchers. The infrastructure includes the following L1 and L2 learner corpora:

Kolipsi 1 **size: ca. 600 000 tokens**

The Kolipsi 1 corpus [24] contains German and Italian L2 productions including a number of reference texts from L1 writers of the same language. The writers were upper secondary school students, whose native language was usually the respective other language, i.e. German when the text was written in Italian and vice versa. Students performed two writing tasks that consisted of an e-mail to a friend describing a picture story (narrative text genre) and in writing a letter to a friend on holiday planning (argumentative text genre). All data was collected in the school year 2007/2008. 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 normalised 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, origin, socio-economic background and L1 of the learners.

Kolipsi 2 **size: ca. 600 000 tokens**

The Kolipsi 2 corpus [25] was built analogous to the Kolipsi 1 corpus, repeating the study done for Kolipsi 1 seven years later (school year 2014/2015). Design and setup of the study as well as processing of the data stayed the same. However, one of the writing tasks (argumentative text) changed slightly.

KoKo **size: ca. 800 000 tokens**

KoKo [4, 26] is a German L1 corpus of argumentative student essays collected from different German speaking regions to illustrate pluricentric language varieties. The total of 1503 texts was elicited during school classes in the year 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 manually for surface features such as student corrections or graphical elements as well as orthographic errors including the assignment of a normalised 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.

Merlin **size: ca. 280 000 tokens**

Merlin [27] is a trilingual learner corpus illustrating European reference levels. The corpus contains 2,290 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 in 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.

One School many Languages **size: ca. 240 000 tokens**

The data collected during the project One School many Languages [28] represents a longitudinal corpus of German, Italian and English texts, composed by the same middle school students during the course of 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. [29]). The dataset that was designed to allow the analysis of multilingual competences contains 1265 picture stories and 1265 opinion texts written in hand-written format 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 normalised form (i.e. target hypothesis) and automatically for sentence splitting, tokenization, lemmatization and part-of-speech tagging.

While the five corpora comprise L1 and L2 texts, respectively, for various languages and language backgrounds of the authors, all texts were elicited in educational settings, were produced in hand-written format and digitized and processed later on through manual transcription. None of the corpora contained multimodal or spoken language.

By making these corpora available, adhering to the FAIR principles for data stewardship, we aim to make the data Findable, Accessible, Interoperable and Reusable for the greater academic public, giving further value to the laboriously created data. In the following we describe the necessary steps we devised for providing our corpora according to the FAIR principles, listing our solutions to the requirements defined by Wilkinson et al. [19].

3.1 Findability

Findability is the first and most basic FAIR principle. Without being able to find a specific corpus or learn about its existence through search interfaces, the data could as well not exist at all. Within the creation of our LCI, Findability is being achieved by depositing all data in the institute's research data repository, Eurac Research CLARIN Centre (ERCC)[9]. The repository uses the CLARIN DSpace software[10] which has been developed by the Institute of Formal and Applied Linguistics, Charles University Prague within the LINDAT/CLARIN project and has been adopted and refined by many other members. 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).

The FAIR principles also state that data are described with rich metadata (F2). Within the Learner Corpus Infrastructure, 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 and 3.4).

Moreover, because the ERCC is part of the European CLARIN infrastructure[11] [30], every item that is being deposited will have its metadata automatically provided via OAI-PMH[12] in various formats and this in turn is periodically being harvested by search engines like the CLARIN Virtual Language Observatory (VLO) and the OLAC catalogue[13]. 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[14].

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 learner corpora available through the ERCC, it is ensured that (meta)data are retrievable by their identifier using a standardised 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 authorisation 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[15] 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 honoured, there are regulations in place that even if data has

[9] https://clarin.eurac.edu
[10] https://github.com/ufal/clarin-dspace
[11] https://www.clarin.eu
[12] https://www.openarchives.org/pmh/
[13] http://search.language-archives.org/
[14] https://www.clarin.eu/resource-families/L2-corpora
[15] https://www.clarin.eu/node/3788

to be removed for whatever reason, the metadata will stay online, including a note on when and why the data was removed.[16]

3.3 Interoperability

In order to achieve Interoperability, the FAIR guiding principles for data stewardship suggest to use 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 learner corpus research - hence not broadly applicable. Moreover, metadata was originally not saved in structured, machine-actionable formats but saved in spreadsheets, XML headers or tab-separated text files. The used vocabularies were scarcely documented and had usually just a single project lifetime.

In order 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, the language for knowledge representation used when offering corpus data in the ANNIS corpus query software[17] [31]. By offering this format, users can make use of the Salt and Pepper conversion framework [32] to convert the data into many other formats used for corpus linguistics in general and learner corpus research in particular.

The metadata for the whole corpus will be provided using the component metadata infrastructure (CMDI) format[18] [33], a machine-actionable metadata format developed within the CLARIN community and supported by the 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. [12]) is considered for future times, especially because it allows to include 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

[16] See also the corresponding point in the ERCC FAQ: https://clarin.eurac.edu/repository/xmlui/page/faq#how-to-delete
[17] https://corpus-tools.org/annis/
[18] https://clarin.eu/cmdi

a standardised set of core metadata fields for learner corpora that is based on the suggestions made by Granger and Paquot [34]. Further research will deal with the challenge on how to make these vocabularies findable, accessible, interoperable and reusable as well. This is currently discussed within the CLARIN community in a specialised taskforce that one of the authors of this paper is a member of.

Finally, in order to include qualified references to other metadata and data, all data are described on one central platform[19] 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.

3.4 Reusability

The final and hardest step, however, is to provide reusable data. While reusable data has to 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) [19].

While 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 future infrastructure should externalise this knowledge in a standardised 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 [34]. While 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 [34]).

The administrative metadata comprises 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 was the data 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.

Finally, 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[20]). The possibility to redistribute the data using this license was provided by former actions ensuring legal and ethical use of the data, i.e. the acquisition of an explicit user consent of the writers or their parents in case they were still minors, the

[19] https://www.porta.eurac.edu/

[20] The license can be found in the respective DSpace collection for each corpus, e.g. https://gitlab.inf.unibz.it/commul/koko/data/bundle/blob/master/EULA-CLARIN-ACA-BY-NC-NORED.pdf. The license text is identical for each corpus except for the corpus name.

anonymisation of the textual data and pseudonymisation of the metadata and the preparation of workflows to allow writers to make use of the right to withdraw their data usage consent or to be informed about what data that is stored and what it was used for, as announced in the respective privacy policy and usage consent forms.

4 Conclusion and outlook

In this article, we have defined the technical and theoretical needs for an infrastructure that is able to make the learner corpora existing at the Institute for Applied Linguistics at Eurac Research available while following the guidelines for FAIR research data. Although our infrastructure is still in its construction phase, especially the requirements for the use of FAIR vocabularies and the data documentation are still in progress, we have implemented most of the requirements established before. Previously published datasets like the Merlin corpus and the KoKo corpus are already available through the ERCC research data repository and additionally through a corpus query interface based on ANNIS and will be updated to match the established requirements for making them FAIR. The other corpora are currently being evaluated and cleaned on various levels and in the process of receiving the necessary documentation and treatment for being findable, accessible, interoperable and reusable through the newly created infrastructure as well. In addition, we plan further research and international collaboration on the technical integration and FAIRification of core metadata definitions as well as on technical possibilities to FAIRify the used vocabularies in general.

We believe that for any new learner corpus project that entails data collection, considering community-specific and domain-relevant research data management right from the start undoubtedly adds substantial value to the created resource and helps avoiding unnecessary theoretical considerations as well as manual work, incomparable corpus resources, stand-alone research results with no means for replication or uptake by others as well as single-use data or simply data that cannot be made available at all because of legal or ethical constraints. While it is possible for new data collection projects to leverage their work substantially by building on existing infrastructures, utilizing technical and/or theoretical efforts already made, existing data often does not allow to "re-build" the data source from the start. Design choices already made often leave little space to adapt to the requirements for reusable resources. Furthermore, metadata that has not been collected during the data collection, is practically impossible to retrieve at a later stage [35].

In this article, we described what we think would be needed to make learner corpora findable, accessible, interoperable and reusable technically and theoretically and addressed how we plan to adapt existing resources to this scheme. However, in some respects our solutions diverged from the optimal designs, as, whether for theoretical reasons or at least for the time being, major redefinitions of already existing data were not feasible. We tried to address this issue by giving pointers to recommendations made by the community, while describing our own solutions next to it. We hope the considerations illustrated in this article will help other researchers to learn from our experience and motivate more people in learner corpus research to share and align their resources with the community.

In our future work we plan to develop best practice guides based on the experience and achievements with our learner corpora infrastructure presented in this paper, with which we aim to support fellow researchers in the learner corpora community who are tasked to compile new learner corpora, thereby streamlining the entire process on the level of the whole community and maximizing synergies and potential for reuse. What is more, we also plan to survey the community in order to gain a better insight into the level of information and skills that the community is equipped with in order to be able to adopt the FAIR scenario (for both curation of existing learner corpora as well as for any new born-FAIR resources), and provide training materials and opportunities that will be filling the identified gaps.

References

1. S. Granger, Language and Technology. Encyclopedia of Language and Education. 3rd edition. pp. 1-14. (2017)
2. S. Granger, Computer learner corpora, second language acquisition and foreign language teaching, pp. 3–33. (2002)
3. Centre for English Corpus Linguistics, Université catholique de Louvain, Learner Corpora around the World, https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (2020)
4. A. Abel, A. Glaznieks, L. Nicolas, E. W. Stemle, *LREC'14 Proceedings*, pp. 2414–2421 (2014)
5. R. Laarmann-Quante, K. Ortmann, A. Ehlert, M. Vogel, S. Dipper, *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 444–456 (2017)
6. A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, G.Venturi, *Proceedings of the Tenth International Conference on Language Resources and Evaluation,* pp. 88–95 (2016)
7. A. Glaznieks, A. Abel, V. Lyding, L. Nicolas, E.W. Stemle, Apples - Journal of Applied Language Studies, 8 (3), pp. 5–20 (2014)
8. A. Abel, A. Glaznieks, Dimensionen des Deutschen in Österreich. Variation und Varietäten im sozialen Kontext, pp. 257–282 (2015)
9. N. Nesselhauf, How to use corpora in language teaching, 12, pp. 125–156 (2004)
10. G. Gilquin, S. Granger, The Cambridge handbook of learner corpus research, 3 (1), pp. 9–34 (2015)
11. J. Lenardič, T. Lindström Tiedemann, D. Fišer. Overview of L2 corpora and resources, https://office.clarin.eu/v/CE-2018-1202-L2-corpora-report-ver2.pdf (2018).
12. E.W. Stemle, A. Boyd, M. Janssen, A. Rosen, D. Rosén, E. Volodina, Selected Papers from the Fourth Learner Corpus Research Conference 2017. pp. 437–478 (2019)
13. E. Volodina, M. Janssen, T. Lindström Tiedemann, N. M. Preradović, S. Ragnhildstveit, K. Tenfjord, K. de Smedt, CLARIN Annual Conference 2018, p. 86–90 (2018)
14. T. Lindström Tiedemann, J. Lenardič, D. Fišer, CLARIN Annual Conference 2018, p. 142-146 (2018)
15. M. Paquot, L. Plonsky, International Journal of Learner Corpus Research, 3 (1), pp. 61–94 (2017)
16. S. Gries, Journal of Second Language Studies, 1 (2), pp. 276–308 (2018)
17. S. Granger, M. Paquot, CLARIN Workshop on Interoperability of Second Language Resources and Tools, https://sweclarin.se/swe/ workshop-interoperability-l2-resources-and-tools (2017)
18. E. Volodina, B. Megyesi, M. Wirén, L. Granstedt, J. Prentice, M. Reichenberg, G. Sundberg, SLTC 2016-The Sixth Swedish Language Technology Conference (2016)
19. M.D. Wilkinson, M. Dumontier, Ij. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P.E. Bourne, Scientific data, 3. (2016)
20. D. Van Uytvanck, H. Stehouwer, L. Lampen, *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pp. 1029-1034 (2012).

21. B. Megyesi, L. Granstedt, S. Johansson, J. Prentice, D. Rosén, C.-J. Schenström, G. Sundberg, M. Wirén, E. Volodina, *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL 2018)*, pp. 47–56 (2018)
22. C. Chiarcos, Linked Data in Linguistics, pp. 161–179 (2012)
23. E. Volodina, L. Granstedt, B. Megyesi, J. Prentice, D. Rosén, C.-J. Schenström, G. Sundberg, M. Wirén, Swedish Language Technology Conference (2018)
24. A. Abel, C. Vettori, K. Wisniewski, KOLIPSI: gli studenti altoatesini e la seconda lingua; indagine linguistica e psicosociale= KOLIPSI: die Südtiroler SchülerInnen und die Zweitsprache; eine linguistische und sozialpsychologische Untersuchung (2012)
25. C. Vettori, A. Abel, KOLIPSI II: gli studenti altoatesini e la seconda lingua; indagine linguistica e psicosociale= KOLIPSI II: die Südtiroler SchülerInnen und die Zweitsprache; eine linguistische und sozialpsychologische Untersuchung (2017)
26. A. Abel, A. Glaznieks, L. Nicolas, E. W. Stemle, *CLIC-it 2016 Proceedings* (2016)
27. A. Boyd, J. Hana, L. Nicolas, D. Meurers, K. Wisniewski, A. Abel, K. Schöne, B. Štindlová, C. Vettori, LREC'14 Proceedings, pp. 1281–1288 (2014)
28. M. Stopfner, Plurilinguismo: nuovi scenari teorici e didattici, Lend, lingua e nuova didattica (special issue) (forthcoming)
29. L. Zanasi, M. Stopfner, La didattica delle lingue nel nuovo millennio. Le sfide dell'internazionalizzazione, pp. 135–148 (2018)
30. E. Hinrichs, S. Krauwer, *LREC'14 Proceedings*, 1525–1531 (2014)
31. T. Krause, A. Zeldes, Digital Scholarship in the Humanities, 31 (1) (2016)
32. S. Druskat, V. Gast, T. Krause, F. Zipser, *Proceedings of LREC'16*, pp. 4492–4499 (2016)
33. D. Broeder, M. Windhouwer, D. Van Uytvanck, T. Goosen, T. Trippel, Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme 1, pp. 1–4(2012)
34. S. Granger, M. Paquot, Core metadata for learner corpora. Draft. (2017)
35. F. Barker, A. Salamoura, N. Saville, *The Cambridge handbook of learner corpus research*, pp. 511–533 (2015)