The Pipeline for Publishing Resources in the Language Bank of Finland

Ute Dieckmann ute.dieckmann@helsinki.fi
Mietta Lennes mietta.lennes@helsinki.fi
Jussi Piitulainen jussi.piitulainen@helsinki.fi
Jyrki Niemi jyrki.niemi@helsinki.fi
Erik Axelson erik.axelson@helsinki.fi
Tommi Jauhiainen tommi.jauhiainen@helsinki.fi
Krister Lindén krister.linden@helsinki.fi

Department of Digital Humanities
University of Helsinki, Finland

Abstract

We present the process of publishing resources in Kielipankki, the Language Bank of Finland. Our pipeline includes all the steps that are needed to publish a resource: from finding and receiving the original data until making the data available via different platforms, e.g., the Korp concordance tool or the download service. Our goal is to standardize the publishing process by creating an ordered checklist of tasks with the corresponding documentation and by developing conversion scripts and processing tools that can be shared and applied on different resources.

1 Introduction

The Language Bank of Finland (Kielipankki, "The Language Bank") is a collection of services for researchers using language resources in digital humanities and social sciences. The Language Bank is coordinated by FIN-CLARIN, a Finnish consortium of universities and research organizations. The general goal of the Language Bank is to make corpora and related tools available to users. Various types of resources can be deposited in the Language Bank, including text and speech corpora, lexicons and terminologies, and many kinds of data sets produced by research projects.

The Language Bank supports public, academic as well as restricted license categories and offers multiple services for providing access to different resource variants. Since the publication framework is complex and not yet sufficiently automatic, depositors cannot upload their resources to the Language Bank and publish them there directly. However, the Language Bank helps and supports the depositors in clearing the licenses and in converting, annotating and describing their data. Thus, unlike other CLARIN centres, the Language Bank participates to some extent in most of the steps in the process where a researcher or a research group deposits a resource with the Language Bank for redistribution. There are both advantages and disadvantages to this approach, which will be discussed in this article.

Ideally, all resources published via CLARIN services should meet the FAIR standards: they should be findable, accessible, interoperable as well as reusable[1]. By creating a shared and well-documented workflow and by using common tools, we aim to ensure that all resources and their future versions are processed, published and maintained in a consistent, transparent and interoperable way.

2 The Publication Framework

An overview of the publication framework of the Language Bank is shown in figure 1. The process is started by entering the new resource to the publishing pipeline (cf., the left column of the figure). The

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/

[1] https://www.clarin.eu/fair

Ute Dieckmann, Mietta Lennes, Jussi Piitulainen, Jyrki Niemi, Erik Axelson, Tommi Jauhiainen and Krister Lindén 2023. The Pipeline for Publishing Resources in the Language Bank of Finland. *Selected papers from the CLARIN Annual Conference 2022. Ed. by Tomaž Erjavec and Maria Eskevich.* Linköping Electronic Conference Proceedings 198, pp. 33–43. DOI: https://doi.org/10.3384/ecp198004

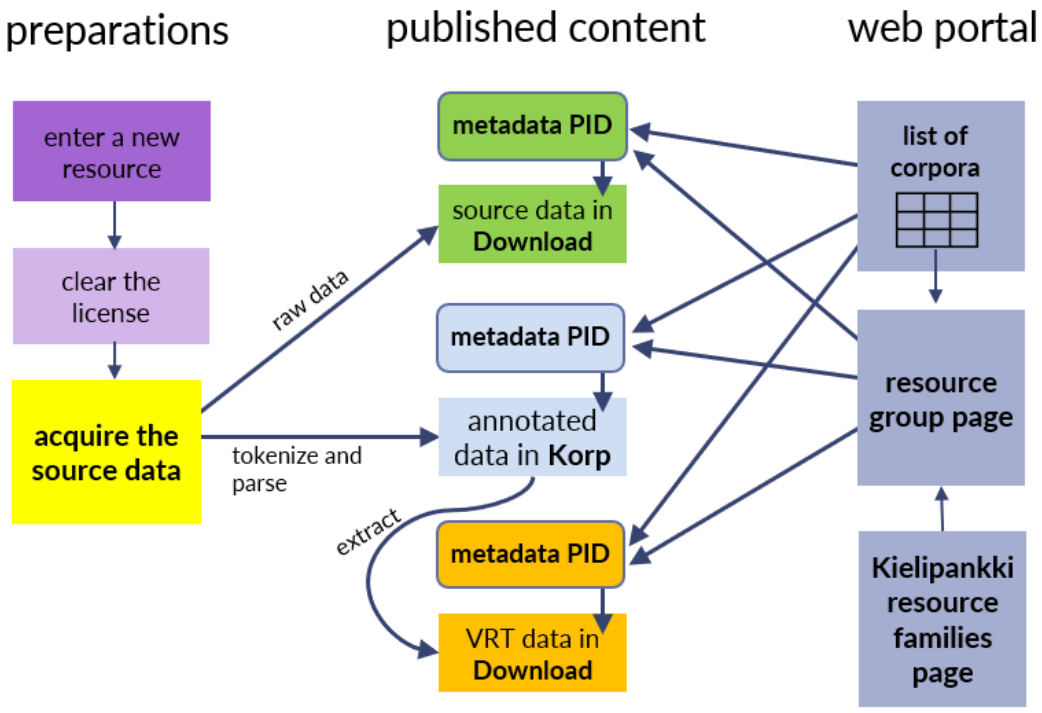

Figure 1. The structure of the resource publication framework

data is then described and prepared for the different means and formats of publication (the middle column). In order to make it easy for users to locate and to use the corpora they need, various pages are created on the website of the Language Bank ("the portal") with additional information about the available resources (the right side of the figure).

2.1 Means of Publication

The resources published in the Language Bank of Finland may be available via the online concordance tool Korp, developed by Språkbanken, the Swedish Language Bank (see Borin et al., 2012), and adapted for the Language Bank of Finland[2]. Korp is a web-based tool that allows the user to search for keywords and more complex constructs in text corpora that are typically enriched with grammatical and other types of annotations. In Korp, it is possible to generate concordances, to compute statistics based on various attributes in the corpora, and to download the results. Korp is well supported and it is used in several CLARIN centres.

Resources can also be downloaded via the download service[3] of the Language Bank. In the download service, we usually provide the original source data as well as the converted data in VRT (*VeRticalized Text*) format as extracted from Korp. Several variants can be offered of the same resource. For the users' convenience, copies of selected versions of the downloadable corpora are also accessible in the computing environment at CSC – IT Center for Science[4]. This makes it easier for users to process the content since they do not have to download and unzip large packages, and the required software can be made readily available on the server.

Lexical resources can be made available via Sanat[5], a WikiMedia-based platform. However, Sanat is currently not part of our official pipeline, since the content is maintained in a community-driven fashion by individual research groups.

The Korp version in the Language Bank supports video and audio playback to a limited extent. In case a speech corpus includes transcripts of the original recordings, the transcripts can be used to create

[2] https://www.kielipankki.fi/korp/
[3] https://www.kielipankki.fi/download/
[4] https://www.csc.fi
[5] https://sanat.csc.fi

a text version of the corpus. The texts can be equipped with links to the corresponding media files, helping the user to locate the original recordings when needed. In case the transcribed text has been manually or automatically aligned with the original recordings, the words and sentences in the transcripts can be annotated with time stamps. The timing information then allows the Korp interface to play the corresponding portion of the media to the user when requested. Korp does not include features for analysing speech signals or video content. However, Korp can be used to provide basic access to transcribed speech corpora.

For storing the internal backup copies of each resource, we use IDA[6], a research data storage service organized by the Ministry of Education and Culture in Finland. IDA is offered free of charge to Finnish universities, universities of applied sciences and state research institutes. The service allows researchers and teams not only to save, organize and share their research datasets but also to freeze the data, i.e., to describe datasets and to store them in an immutable state for long-term archiving.

2.2 Access Rights

The Language Bank aims to provide resources as openly as possible. Many resources can be made publicly available (CLARIN PUB license category). However, access restrictions may be necessary in case the resource includes, e.g., copyrighted content or personal data that should be protected.

Some resources are licensed for academic use only (ACA), and they may be accessed by signing in with credentials issued by the user's home institution. Furthermore, the Language Bank is able to distribute resources under restricted licenses (RES), in which case users can apply for individual access rights in the Language Bank Rights (LBR) service[7]. LBR currently supports federated login and user identities via CLARIN[8] or Eduuni[9]. The Language Bank uses the common CLARIN licensing framework, with some local adjustments[10]. Unless the original material has been previously available under a public license, the licenses of individual resources in the Language Bank are based on agreements with the rightholders and, in the case of resources that contain personal data, with the data controllers.

In some cases, it is possible to offer several variants of the same resource under different licenses. For instance, since speakers might be identifiable based on their voice, audio speech recordings often need to be protected, e.g., by restricting access to them. However, it may be possible to make the anonymized or pseudonymized transcripts available under a less restricted license for specific purposes where access to the audio is not needed.

The Language Bank Rights system is based on REMS (Resource Entitlement Management System)[11], an open-source electronic tool developed by CSC for the management of access rights to research data. Researchers can log into the system by using the user credentials provided by their home organisation. The researcher selects a resource for which access rights are applied, fills in an electronic application form, and agrees to the terms of use for the dataset in question. The application can then be circulated via LBR to the rightholder's representative for approval. LBR can also be used, e.g., for contacting the supervisor of a student applicant in case their endorsement is required before access can be granted to a specific resource. From LBR, it is also possible to obtain reports on applications and approved access rights.

2.3 Documentation

While developing our publishing processes, we have paid attention to the systematic documentation of the resources as well as to improving their findability. Corpora can be found either on the list of published resources or on the list of forthcoming resources in the portal.

Over time, several versions and variants have been published of individual resources on different platforms. To help the users find out which version is the most relevant one for them, we use the so-called resource group pages for documenting all the versions and variants of a given resource as a group. The resource group pages may include additional resource-specific instructions that cannot be included

[6] https://ida.fairdata.fi
[7] https://lbr.csc.fi
[8] https://www.clarin.eu/content/clarin-identity-provider
[9] https://info.eduuni.fi/en/services/eduuni-id
[10] https://www.kielipankki.fi/support/clarin-eula/
[11] https://www.csc.fi/rems-kayttovaltuuksien-hallintajarjestelma

in the individual metadata records (especially when there are many versions). The resource group page of a given resource lists all available versions of the resource, including links to their metadata records, their access locations, and further information. A link to the resource group page can be found in the metadata record of each version of the resource.

Following the example of CLARIN Resource Families[12], we also offer a portal page where the resource groups in the Language Bank are categorized under CLARIN-style families.

3 Challenges and Goals

While implementing and developing our publishing pipeline, we aim to meet the needs of the users as well as to improve our internal workflows. Resources should be offered in consistent and interoperable formats and they should be easy to find and to process by researchers and research groups.

More than 250 resources are currently available via the Language Bank of Finland. About 100 resources are listed as forthcoming, and more are added every month. Before implementing the publishing pipeline, each team member involved in the process of publishing resources had their own workflows and scripts for converting data. This often resulted in slight inconsistencies in the published resources. In addition, it was not easy to monitor the state of each resource within the publication process. Certain tasks, such as parsing the data for publication in Korp, were carried out by only one person in the team, making processes very dependent on this person's availability and time.

The process of publishing an individual corpus usually involves 3–4 people in the Language Bank. In case of an exceptionally simple and well-described dataset with no licensing issues, it does not take more than one or two working days to publish the source data for download. If intense license discussions and several different means of publication are required, the process can take up to 60 working days.

Our aim is to perform faster. Ideally, the process should enable us to publish resources within a shorter time frame and to reduce the amount of time that a certain resource ends up waiting in the pipeline for the next processing steps. To make the workflow more efficient, it is important to be able to monitor the status of a resource during the publishing process and to share tasks and knowledge within the team. For this purpose, we have collected a list of the specific tasks that are addressed during the publication process of most resources.

4 Tasks within the Publishing Pipeline

For each new resource, we maintain a checklist[13] of the tasks in the shared pipeline that are relevant for the resource in question. The list is used for keeping track of the status of the resource during the publishing process. Some tasks on the list are mandatory for all types of resources, whereas others are applicable to specific types only. According to the type of the task, which can be for example administrative or technical, work can be assigned to a person with the required skills.

The tasks can be classified into different groups, as shown in figure 2, starting at entering the new resource and clearing the license, and finishing at publishing the resource in the different channels and formats. Grouping the tasks thematically as well as chronologically helps us keep track of the status of each resource and to make the checklist of tasks clearly laid out. It is to be noted, however, that the order of the events in the publication pipeline is not a strict timeline but rather a dependency structure. For instance, data acquisition and processing may in some cases start before the deposition license agreement with the data provider is signed, whereas the corpus cannot be published unless the license is cleared.

A ticketing system helps in managing and monitoring the publication processes of individual resources. For this purpose, we use Atlassian JIRA. Currently we are looking for a more convenient way

[12] https://www.clarin.eu/resource-families
[13] http://urn.fi/urn:nbn:fi:lb-2023032703

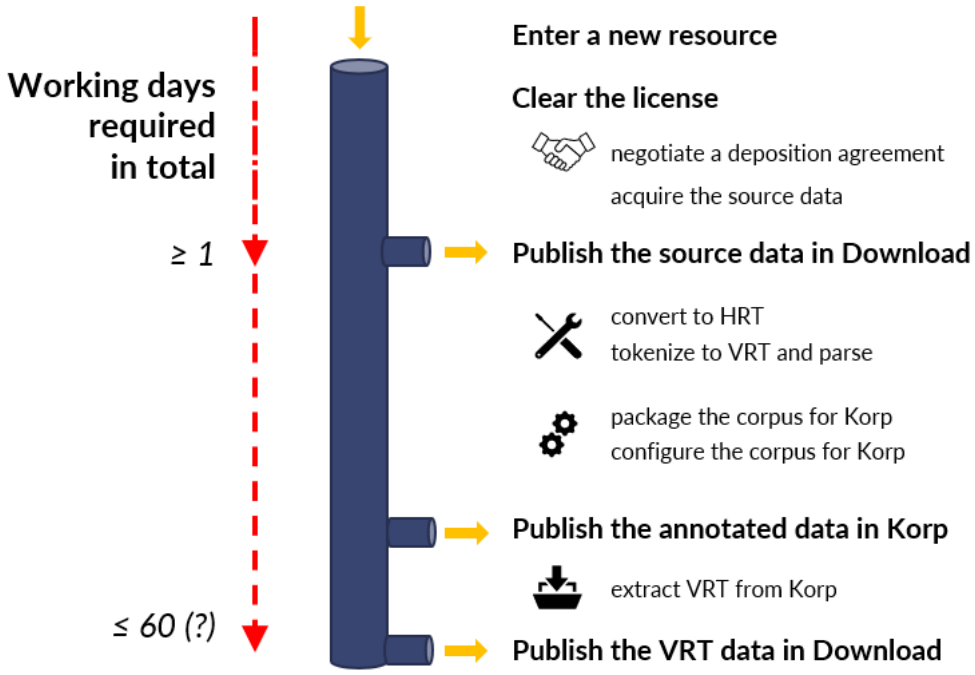

Figure 2. Tasks within the publishing pipeline

of implementing the list of tasks to the pipeline, so that tasks could be automatically created for each incoming resource and made more accessible to the team.

4.1 Entering a New Resource to the Pipeline of the Language Bank of Finland

When a researcher or a research group creates a new resource that they wish to make available to other researchers or publicly, they are first asked to submit the most important details regarding the resource by filling in an e-form[14]. The Language Bank then creates a preliminary metadata record on the local META-SHARE repository[15], where the metadata of all resources available via the Language Bank are currently maintained. The preliminary metadata are checked together with the depositor. The details can be updated and amended later. In order to make sure that the metadata records meet our quality standards and remain consistent, the editing rights to the metadata records are restricted to a few people with the required expertise.

The metadata records in the META-SHARE node of the Language Bank are harvested by our OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) metadata provider that transforms the native META-SHARE records to CLARIN's CMDI format. The metadata are then available to external services, such as the Virtual Language Observatory[16] maintained by CLARIN. Our local provider validates the records before and after the transformation, using the corresponding XML schemas, to avoid providing malformed metadata and to inform our metadata maintainers about faulty records. The XSLT script transforming the data to the CMDI format was provided by the implementors of the META-SHARE format.

[14] http://urn.fi/urn:nbn:fi:lb-2021121422
[15] https://metashare.csc.fi
[16] https://vlo.clarin.eu

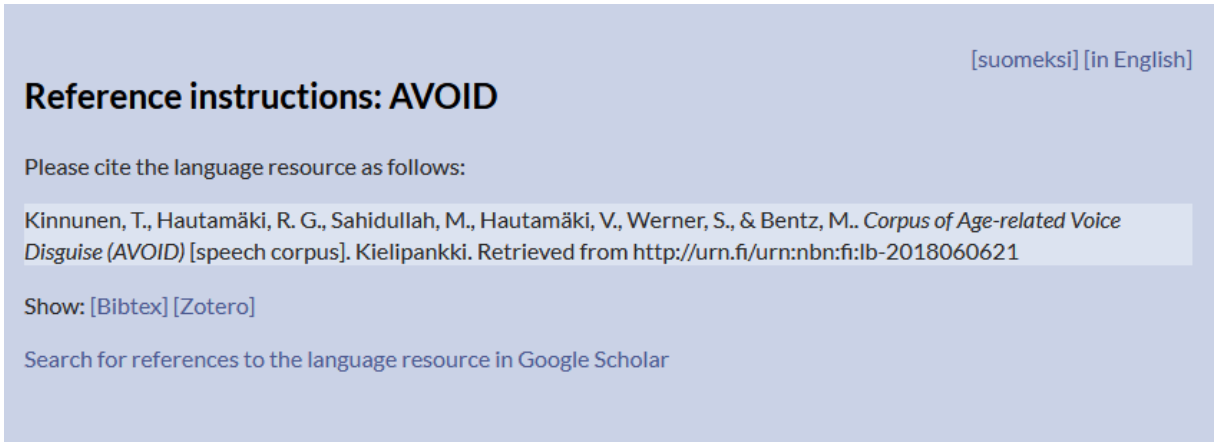

Figure 3. An example of reference instructions

For publications, researchers may need a persistent reference to their resource before it is made available by the Language Bank. Since unofficial links should be avoided in citations, the Language Bank assigns a persistent identifier (PID) to the metadata record as soon as the resource exists and has been sufficiently well described. Each version of the resource gets a metadata record and a PID of its own. For instance, the downloadable source version will have a different PID from the Korp version of the resource. The metadata PID is the citable and primary identifier of the resource version. For internal use, we also assign PIDs to the access location, to the license pages and to the resource group pages. The Language Bank uses the URN system for generating PIDs. For details on how the Language Bank uses PIDs, see Matthiesen and Dieckmann, 2019.

At this point, the resource is added to the list of forthcoming resources on the website of the Language Bank. The reference instructions will then be automatically generated and displayed on demand. A link to the reference instruction is included in the META-SHARE record of the resource and in the resource group page. The reference instruction (for an example, see figure 3) currently provides the user of a resource with the names of the author(s) listed in the order required by the rightholder in the deposition agreement, the publication year of the dataset, the full title of the corpus, the type of the dataset (e.g., [text corpus]), the name of the centre or repository that is maintaining the resource for access (The Language Bank), and, lastly, the citable PID of the resource in question.

In addition to the full title, each version of a resource is also labelled with a short name that is included in the corresponding metadata record. The short name typically consists of a stem, derived from the official name or acronym of the corpus, an optional version number, and a suffix that carries information about the type of the resource variant (cf. figure 4). The short names can be used as keys for keeping the resource versions distinct from each other within various services, e.g., in the names of portal pages and in the file and folder names of downloadable corpora. The short names are also used as keys for generating the citation instructions in the web portal. As shown in the examples in figure 4, the members

Abbreviation	Name and metadata
wanca2016-korp	Wanca 2016, Korp Version
wanca2016-src	Wanca 2016, source
wanca2016-vrt	Wanca 2016, VRT

Figure 4. Examples of the short names of three resource variants within the resource group 'wanca'

of the Wanca resource group can be conveniently kept together since the stem ‘wanca’ is systematically included in the short names of the different versions and variants of the resource.

4.2 Clearing the License for the Resource and Acquiring the Source Data

Unless the resource has previously been published under an open license, the Language Bank and the depositor negotiate on the license for distributing the resource. In case there are no legal reasons for restricting the use of the resource to specific purposes, users or user groups, it is preferable to select an open license. However, language resources may contain for instance copyrighted text, and it is necessary to ensure that the depositor has sufficient rights to share the material via the Language Bank. Moreover, the resource (or at least certain parts of it) may include personal data, which must be considered so that the appropriate safeguards can be applied when the material is stored and processed by the Language Bank and by the end-users in their own research projects.

In case the resource contains copyrighted content, additional steps may be needed to obtain permissions from the copyright holders. These permissions are usually requested by the researcher who wishes to deposit the resource, but the Language Bank can offer support for formulating the requests.

If the resource includes personal data, the data controller, responsible for the original purpose of processing the data, is involved in the deposition agreement. In this case, the end-user license will include the condition +PRIV, and all users who access the resource via the Language Bank will be required to comply with the resource-specific data protection terms and conditions.

The Language Bank uses a generic deposition license agreement template[17]. In order to discuss the details, a meeting with the depositor is often needed. When an agreement is reached, the end-user license is published in the portal. Using PIDs, the metadata record will refer to the license page and vice versa.

After receiving the source data from the resource depositor, the data are checked for format and validity, and a description of the contents is added for internal use. A backup copy of the data is stored in IDA.

4.3 Publishing the Source Data in Download

Since the data conversion process tends to take time, the first version of a corpus to be published in the Language Bank is usually the source data that is made available for download. In this version, the original content is not modified. However, the metadata and license information must be available and up to date.

A PID is added to the metadata record, and a resource group page, which also gets a PID, is created and linked with the corresponding metadata. The source version of the resource, and possibly some other foreseen versions of the resource, are added to the list of forthcoming resources in the portal, to keep the corpus owners and the potentially interested researchers informed. The attribution details are added to the metadata record. Furthermore, PIDs are assigned to the license page and the download location.

[17] https://www.kielipankki.fi/support/dela/

To prepare the resource for download, the source data is packaged into one or more zip files as agreed with the corpus depositor, including a README text file that contains basic information on the resource and a LICENSE text file offering information on the access rights for this resource. In case the license of the resource is in the RES category, a record is created on the LBR system in order to be able to control access to the download location. Similarly, if the license is ACA, academic user login will be required to download the resource. After a successful check of the quality and accessibility of the uploaded zip packages, the metadata record and the resource group page can be updated with the access location PID.

To finalize the publishing of a resource, it is moved from the list of forthcoming resources to the list of published resources in the portal. A news item is published in the portal to inform interested researchers about the new resource. The depositor is informed about the publication as well. The download package is uploaded to be stored in IDA, and in selected cases, for example when wide use of a large resource is anticipated, the unpacked source data is also made available in CSC's computing environment.

4.4 Publishing the Data in Korp

Our goal is to make data accessible to the user in a uniform format, converted from various source formats. For this purpose, we use VRT, which is the input format for the IMS Open Corpus Workbench (CWB) software (Evert and Hardie, 2011) underlying Korp. The data is first tokenized, and annotations are inserted in order to include any descriptive information available in the source data, such as the dates, locations and authors of the individual texts. The data is then lemmatized, tagged with parts of speech and/or parsed, depending on the automatic annotation tools available for the language in question. The data can also be extended with additional annotations, such as name annotations, sentiment annotations and identified languages.

The first steps of publishing the Korp version of the resource are similar to those of the downloadable versions. A metadata record for the Korp version is created or updated and PIDs are assigned to the metadata record and to the access location. In case the license of the resource is RES, an LBR record is created.

The format of the original data tends to vary between corpora. It can be for example plain text, PDF, RTF, or an XML format such as TEI. For Korp, PDF documents are first converted into text files. Unless the source data is already tokenized, the first aim is to convert this data to a simple, XML-style format which must be UTF-8-encoded Unicode. An example of this format, which we call HRT (*HoRizontal Text*), is shown in figure 5. The conversion is carried out with tailor-made scripts and it can often be the most time-consuming step, depending on the format of the original data. The basic idea is to segment the content of the original files so that the plain text is inside text and paragraph tags, which can include descriptive attributes. These files with a relatively simple structure are then used as input for further processing tools. The next step is the tokenizing process where the paragraphs are segmented into sentence elements and tokens. The output format of the tokenizer is VRT, which we have extended with a comment that provides names for the otherwise positional attributes of tokens. An example of the VRT format can be seen in figure 6.

It is possible for the Language Bank to apply further tools on the VRT data to add any desired annotations, while preserving the sentence and token boundaries and previous annotations. For instance, information about the languages used in the text can be added by running a language identifier such as HeLI-OTS (Jauhiainen and Jauhiainen, 2022) that includes language models for 200 languages.

For Finnish and other languages with a parser and named-entity recognizer available, the parsing process is carried out on the validated VRT data. For years, we have been using an early version of the Turku dependency parser for Finnish, developed by the Turku NLP group and adapted for VRT. We are currently adopting their new neural parser[18] along with the Universal Dependencies annotation model.

A single script calling several other scripts handles the processing of VRT files to create a Korp corpus package containing the CWB data files and the Korp MySQL database import files. The resulting package is then installed on the Korp server, and a Korp corpus configuration is added with the information on the corpus and its annotations (attributes). The corpus configuration determines where and how the corpus is shown in Korp. The configuration is first added to a test instance of Korp. When

[18] http://turkunlp.org/Turku-neural-parser-pipeline/

```
<text binding_id="1377028" date="1986" datefrom="19860101" dateto="19861231">
<paragraph id="0">
mahdollista.
</paragraph>
<paragraph id="1">
Malminkartano on kuuluisa linnastaan, jota eräässä
kulttuurihistoriallisessa asiantuntijalausunnossa on kehuttu Suomen
komeimmaksi kartanorakennukseksi. Se valmistui 1885 entisen puutalon
tilalle, jonka alakerros oli peräisin 1600-luvulta.
</paragraph>
<paragraph id="2">
Arkkitehtina oli F.A. Sjöström ja tyyli on Suomen maaseudulla melko
harvinaista uusrenessanssia, jonka arvostus aleni pian siinä määrin,
että rakentajan oma tyttärentytär puhui ”maun
rappiokaudesta”. Myöhemmin on ymmärtämys sitä kohtaan kasvanut.
</paragraph>
</text>
```

Figure 5. An example of the HRT format

the test instance meets the expectations, the configuration is copied to the production Korp, where the corpus is published as a beta version.

The new corpus is announced in the Korp news desk as well as in the portal. The beta status is removed after two weeks unless requests for changes are received during this period. Finally, a copy of the Korp corpus package is stored in IDA.

Although this approach is time-consuming, it has been designed so as to ensure the consistency and interoperability of the published resources. It is important to preserve as much of the information in the original data as possible, be it structural information or metadata at various structural levels of the data. We are able to reach this goal by using tailor-made scripts for converting the source data to HRT.

All the generic and tailor-made scripts used for processing corpora are published openly on GitHub[19].

4.5 Publishing the VRT data in Download

After publishing a resource in Korp, the VRT data is usually extracted from Korp and published in the download service, in order to provide consistent versions of the data via both channels. The VRT version of a resource is published in the download service in the same way as the source data. For the VRT version, a separate metadata record is created, and the corpus version is added to the resource group page.

We decided to make the VRT versions of the data available for the users, since we believe that VRT can be useful for further processing. The VRT format[20] is simple and human readable and easy to process and to transform. It is a combination of one-word-per-line (vertical) format and simple XML markup. In the future, we also intend to offer tools for easy conversion from VRT to other formats.

4.6 Testing and quality control

When a corpus is ready to be published in a given means of publication, the final step before the actual publishing is quality control, i.e., a testing procedure is required. Ideally, the testing should be carried out by a member of the team not involved in the processing of the resource in question. Testing procedures are tailored for download and Korp separately. They also differ between (versions of) resources with different access rights. The accessibility on the different platforms is tested, and it has to be made sure that ACA and RES restrictions work as expected.

Our testing procedures are still under development. Our aim is to have a catalogue of test cases available, covering what should be tested from the user’s perspective as well as taking internal needs like archiving and documentation into account. A comprehensive checklist covering all the various cases together with clear guidance and possibly screenshots should enable every member of the team to take

[19] http://urn.fi/urn:nbn:fi:lb-2023032701
[20] http://urn.fi/urn:nbn:fi:lb-2023020121

```
<!-- #vrt positional-attributes: ref word lemma pos msd deprel dephead -->
<text binding_id="1377028" date="1986" datefrom="19860101" dateto="19861231">
<paragraph id="0" sum_lang="|xxx:1|">
<sentence id="0" lang="xxx" lang_conf="3.7488587">
1       mahdollista     mahdollinen     A       NUM_Sg|CASE_Par|CMP_Pos ROOT    0
2       .       .       Punct   _       punct   1
</sentence>
</paragraph>
<paragraph id="1" sum_lang="|fin:7|xxx:1|">
<sentence id="1" lang="fin" lang_conf="1.3936844">
1       Malminkartano   malmi|kartano   N       NUM_Sg|CASE_Nom|CASECHANGE_Up   nsubj-cop       3
2       on      olla    V       PRS_Sg3|VOICE_Act|TENSE_Prs|MOOD_Ind    cop     3
3       kuuluisa        kuuluisa        A       NUM_Sg|CASE_Nom|CMP_Pos ROOT    0
4       linnastaan      linna   N       NUM_Sg|CASE_Ela|POSS_Px3        nommod  3
5       ,       ,       Punct   _       punct   11
6       jota    joka    Pron    SUBCAT_Rel|NUM_Sg|CASE_Par      rel     11
7       eräässä eräs    Pron    NUM_Sg|CASE_Ine det     9
8       kulttuurihistoriallisessa       kulttuuri|historiallinen        A       NUM_Sg|CASE_Ine|CMP_Pos amod    9
9       asiantuntijalausunnossa asian|tuntija|lausunto  N       NUM_Sg|CASE_Ine nommod  11
10      on      olla    V       PRS_Sg3|VOICE_Act|TENSE_Prs|MOOD_Ind    auxpass 11
11      kehuttu kehua   V       NUM_Sg|CASE_Nom|VOICE_Pass|PCP_PrfPrc|CMP_Pos   rcmod   4
12      Suomen  Suomi   N       SUBCAT_Prop|NUM_Sg|CASE_Gen|CASECHANGE_Up       poss    13
13      komeimmaksi     komea   A       NUM_Sg|CASE_Tra|CMP_Superl      amod    14
14      kartanorakennukseksi    kartano|rakennus        N       NUM_Sg|CASE_Tra nommod  11
15      .       .       Punct   _       punct   3
</sentence>
```

Figure 6. An example of the VRT format

care of the quality control. Also validation scripts can help, where feasible, as they produce meaningful reports and are usually easy to run.

5 General Discussion

While using and developing further our publishing pipeline for a few years now, we have discovered advantages and disadvantages in our approach. Especially the time and effort required from our side to publish a resource can be seen as problematic in this respect.

After introducing the resource group pages, the resource families page, license pages for public resources, pages for the resource-specific data protection terms and conditions, etc., the number of administrative tasks has increased to some extent. The actual working time spent on publishing one version of a resource might not have changed significantly. However, the waiting time within the pipeline has diminished, since all team members can at least in principle perform the required tasks, due to the shared scripts and documentation. It is now much easier for us to manage the tasks and to monitor the publishing process for each resource. The quality of the published packages and the findability of the corpora has also improved.

We aim to offer resources in uniform formats, to preserve the information in the original data to a maximal extent, to make resources available while respecting the potential restrictions regarding access rights, and to ensure the quality, findability and accessibility of the published resources. These goals may justify the current active role of the Language Bank in the publication workflow. Nevertheless, by developing our processes and by making them more automatic, it will be possible for the resource creators to participate more and more actively in the publication of their data.

The main focus of the publishing pipeline is currently on processing text corpora via Korp and on creating download packages. Ever since the former LAT system was taken out of use in the Language Bank at the end of year 2020, we have been looking for a suitable replacement service that would enable users to query, to browse and to access speech and sign language corpora that may include audio and/or video files as well as time-aligned annotations in multiple annotation tiers. Currently, most of the speech and sign language corpora available in the Language Bank are offered via the download service only, and the users are instructed to use locally installable GUI tools for querying the annotations and for analysing the multimedia content, e.g., ELAN[21] (MPI in Nijmegen) for audio and video, or Praat[22]

[21] https://archive.mpi.nl/tla/elan
[22] http://www.praat.org/

(Boersma and Weenink, 2023) for audio and digital signal processing. It is hoped that new solutions can be developed so as to provide a higher level of service to audio and video corpora in the future.

The Language Bank is currently working towards building a centralized database that includes all the public as well as internal metadata of each resource, their status in the pipeline and the license details. We intend to use the database for generating and maintaining the public metadata records, the README documents included in the download packages, the license pages, the resource group pages, the listings of currently available and forthcoming corpora, etc. The database could also be plugged in with the PID generator and be used to automatically provide the citation instructions. By storing and maintaining all the resource information in one place, we would be able to reduce the need for manually copying and pasting data between a number of documents and web pages. Moreover, the users could be allowed to submit preliminary information directly to the database. The resource database will help us minimize human errors and administrative delays in the resource publication process.

We believe that the different CLARIN centres should be able to collaborate with each other in developing their technical as well as administrative practices. In order to gain a better understanding of what sort of expertise is already available within CLARIN, an overview of the main services, platforms and means of user support that are offered by each CLARIN centre would be useful. We are planning to conduct a survey in collaboration with the Standing Committee for CLARIN Technical Centres and the CLARIN User Involvement Committee.

6 Conclusions

Currently, the Language Bank of Finland provides researchers with access to over 250 resources, and many more are forthcoming. The licensing and publishing process of each resource takes time and effort and requires various kinds of expertise. Based on our experience, we have identified a number of tasks that are relevant when publishing most types of resources, resulting in a checklist and modular documentation[23] offering instructions for the individual tasks. Although this pipeline is still under development, the general workflow has already proven useful for managing and monitoring the publication process more efficiently.

We aim to automate and document our processes even further to enable resource depositors to take a more active role in preparing their data. The pipeline should also be extended in order to make it more convenient for users to discover and to share tools via the Language Bank. We believe that by comparing and sharing good practices with other CLARIN centres, it is possible to support researchers even better.

References

Paul Boersma and David Weenink. (2023). Praat: doing phonetics by computer [Computer program]. Version 6.3.06, retrieved 31 January 2023 from http://www.praat.org/

Lars Borin, Markus Forsberg and Johan Roxendal. (2012). Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012*. Istanbul: ELRA, pages 474–478.

ELAN (Version 6.4) [Computer software]. (2022). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Stefan Evert and Andrew Hardie (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. *Proceedings of the Corpus Linguistics 2011 conference*. University of Birmingham, UK.

Tommi Jauhiainen and Heidi Jauhiainen. (2022). HeLI-OTS 1.3 (1.3). Zenodo. https://doi.org/10.5281/zenodo.6077089

Martin Matthiesen and Ute Dieckmann. (2019). A PID is a Promise – Versioning with Persistent Identifiers. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 103–112.

[23] http://urn.fi/urn:nbn:fi:lb-2023032702