The Resource Publishing Pipeline of the Language Bank of Finland

Dieckmann, Ute

CLARIN ERIC
2022

Dieckmann , U , Lennes , M , Piitulainen , J , Niemi , J , Axelson , E , Jauhiainen , T & Linden , K 2022 , The Resource Publishing Pipeline of the Language Bank of Finland . in T Erjavec & M Eskevich (eds) , CLARIN Annual Conference Proceedings, 2022 . CLARIN Annual Conference Proceeedings , CLARIN ERIC , Utrecht , pp. 53-56 , CLARIN Annual Conference , Prague , Czech Republic , 10/10/2022 .

http://hdl.handle.net/10138/350056

The Resource Publishing Pipeline of 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 check list 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 the collection of services coordinated by FIN-CLARIN, the Finnish consortium of universities and research organizations. Various types of resources can be deposited in the Language Bank, including for instance 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 directly upload their resources. We need to support them in clearing the licenses and in converting, annotating and describing their data.

Currently, more than 250 resources are 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, every team member involved in the process of publishing resources had their own workflows and conversion scripts. This tended to result in slight inconsistencies in the published data. Monitoring the current state of a given resource within the publication process was not always easy. Some tasks, like 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.

Ideally, all resources published via CLARIN services should meet the FAIR standards (findable, accessible, interoperable, 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 Means of Publication

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, Forsberg & Roxendal, 2012),

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

and adapted for the Language Bank of Finland[2]. Resources can also be downloaded via the download service[3] of the Language Bank. 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]. Lexical resources can be made available via Sanat[5]. For storing the internal backup copies of each resource, we use IDA, the research data storage service organized by the Ministry of Education and Culture in Finland[6].

The resource group page of a given resource on the website (the "portal") of the Language Bank 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[7], we also offer a portal page where the resource groups in the Language Bank are categorized under CLARIN-style families.

2.1 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 are often necessary for protecting copyrighted content or personal data. 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[8]. LBR currently supports federated login and user identities via CLARIN[9] or Eduuni[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 right holders 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, anonymized or pseudonymized transcripts could be available under a less restricted license for purposes where audio is not needed.

3 Tasks within the Publishing Pipeline

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.

For each new resource, we maintain a check list[11] 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, whereas some are applicable to specific resources only. According to the type of task, which can be for example administrative or technical, work can be assigned to the person with the required skills.

3.1 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[12]. The Language Bank then creates a preliminary metadata record on the local

[2] https://korp.csc.fi
[3] https://www.kielipankki.fi/download
[4] https://www.csc.fi
[5] https://sanat.csc.fi
[6] https://ida.fairdata.fi
[7] https://www.clarin.eu/resource-families
[8] https://lbr.csc.fi
[9] https://www.clarin.eu/content/clarin-identity-provider
[10] https://info.eduuni.fi/en/services/eduuni-id
[11] https://github.com/CSCfi/Kielipankki-utilities/blob/master/docs/corpus_publishing_tasklist.md
[12] http://urn.fi/urn:nbn:fi:lb-2021121422

META-SHARE repository[13]. The preliminary metadata are checked together with the depositor. The details can be updated and amended later. The metadata are automatically harvested from META-SHARE by other services, e.g., the Virtual Language Observatory[14] maintained by CLARIN.

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. This PID is the citable and primary identifier for the resource (for details on how the Language Bank uses PIDs, see Matthiesen & Dieckmann, 2019). At this point, the resource is also added to the list of forthcoming resources on the website of the Language Bank.

3.2 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. If the resource contains copyrighted content, additional steps may be needed to obtain permissions from the copyright holders. If the resource includes personal data, the data controller 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[15]. 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, they 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.

3.3 Publishing the Source Data in Download

Since the data conversion process tends to take time, the very first version of a corpus to be published 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 other planned versions of the resource, are added to the list of forthcoming resources in the portal, to keep the corpus owners and possibly interested researchers informed. A PID for the download location is requested.

To prepare the resource for download, the source data is packaged into one or more zip files as agreed with the corpus depositor. In case the license of the resource is RES, a record is created on the LBR system in order to control access to the download location. Similarly, if the license is ACA, academic user login will be required to download the resource. After testing the zipped packages, they are uploaded to the download service, together with a README text file containing basic information on the resource and a LICENSE text file offering information on access rights for this resource. The metadata record and the resource group page can then be updated, adding the access location PID.

To finalize the publishing of a resource, it is added 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 IDA, and in selected cases the unpacked source data is also made available in CSC's computing environment.

3.4 Publishing the Data in Korp

If a resource is meant to be made available in Korp, the data is tokenized, parsed (if a parser is available for the language in question), and possibly extended with additional annotations (name annotations, sentiment annotations, identified languages). The first steps of publishing a resource are similar, however, regardless of the means of publication. 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.

[13] https://metashare.csc.fi

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

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

The format of the original data may differ between corpora. It can be for example plain text, PDF, RTF, a tabular format such as CoNLL-X, or an XML format such as TEI. The first aim is to convert this data to a simple form of XML, which must be UTF-8 encoded Unicode. This task is carried out with individually developed scripts and usually is the most time-consuming, depending on the format of the original data. The basic idea is to segment the content of the original files so that 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 (VeRticalized Text), the input format for the IMS Open Corpus Workbench (CWB) software underlying Korp. In addition, we have extended the VRT format with a comment that provides names for the otherwise positional attributes of tokens.

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 could be added by running a language identifier such as HeLI-OTS (Jauhiainen & 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[16] along with the Universal Dependencies annotation model.

One single but complex script handles the processing of VRT files to create a Korp corpus package containing CWB data files and Korp MySQL database import files. The resulting package is then installed on the Korp server and a corpus configuration is added. When the test instance meets expectations, the corpus is published in Korp as a beta test 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 appear during this period. Finally, a copy of the Korp corpus package is stored in IDA.

After publishing a resource in Korp, the VRT data is usually extracted from Korp and published in the download service in order to provide similar 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.

4 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 tends to require 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 check list and modular documentation[17] 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 automatize and document the process even further to enable resource depositors to take a more active role in preparing their data. We believe that by comparing and sharing good practices with other CLARIN centres, it is possible to support researchers even better.

References

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

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.

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

[17] https://github.com/CSCfi/Kielipankki-utilities/tree/master/docs