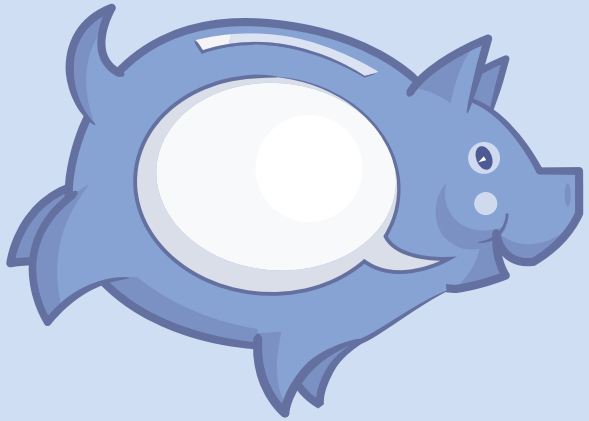


KIELIPANKKI
The Language Bank of Finland

The Resource Publishing Pipeline of the Language Bank of Finland

Ute Dieckmann, Mietta Lennes, Jussi Piitulainen, Jyrki Niemi,
Erik Axelson, Tommi Jauhiainen & Krister Lindén

FIN-CLARIN

https://www.kielipankki.fi

KIELIPANKKI
The Language Bank of Finland

LANGUAGE BANK ACCESS CORPORA TOOLS ORGANIZATION SUPPORT SUOMEKSI PÅ SVENSKA

Access

Apply for rights to use our language resources.

Corpora

Browse our corpora.

Tools

Try our tools.

Organization

Who are the Language Bank?

Support

Help and instructions.

Search the Language Bank Portal:

Researcher of the Month: Filip Ginter

News

Findable

Accessible

FAIR data

Interoperable

Re-usable

FAIR data

Findable

Consistent metadata

Virtual Language Observatory

PID

+ Access location ***PID***

KIELIPANKKI The Language Bank of Finland

www.kielipankki.fi

Accessible

KORP v9

DOWNLOAD

PUB

eduGAIN

ACA

Haka

RES

Language Bank Rights

Interoperable

Instructions for resource creators

HRT / VRT common formats

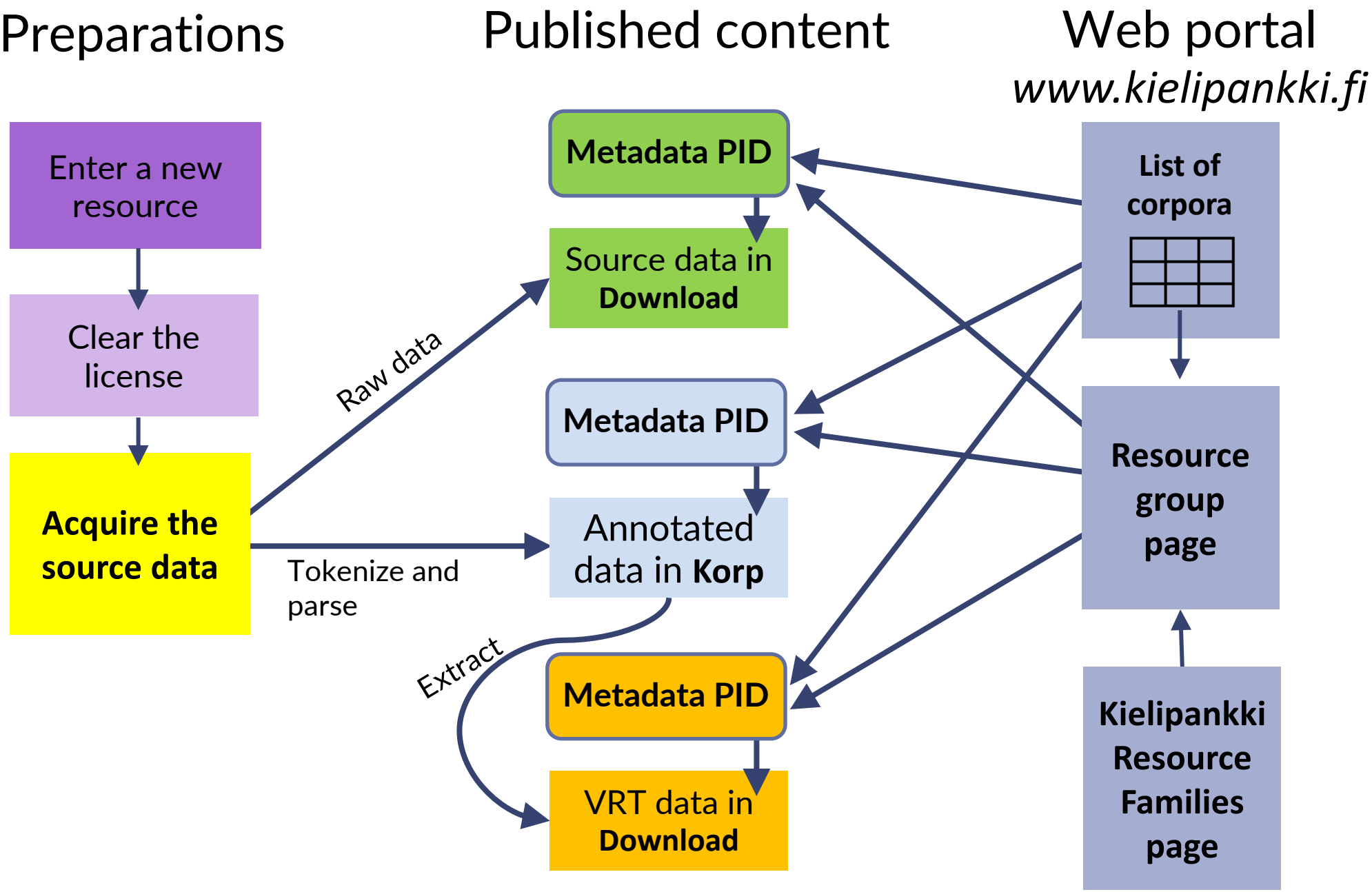
Common processing tools

Re-usable

Deposition agreements

Long-term archiving (?)

Support for versions and variants

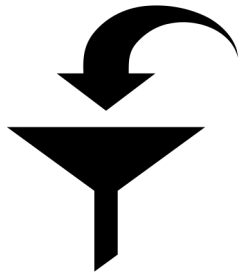

Preparations
Published content
Web portal
www.kielipankki.fi
Enter a new resource
Clear the license
Acquire the source data
Raw data
Tokenize and parse
Extract
Metadata PID
Source data in Download
Metadata PID
Annotated data in Korp
Metadata PID
VRT data in Download
List of corpora
Resource group page
Kielipankki Resource Families page

KIELIPANKKI
The Language Bank of Finland

Challenges

- Keeping resources consistent and interoperable
- Monitoring resource status during the publishing process
- Sharing tasks and knowledge within the team
- Making the workflow more efficient
- Enabling the data provider to take a more active role

FIN-CLARIN



Enter a New Resource to the Pipeline

- Depositor submits resource details in an e-form

- Create a preliminary metadata record
- Request a citable PID for the resource
- Plan and distribute tasks required for publishing

- Please fill in all relevant parts of the form, even if the information provided is still preliminary.
- If necessary, the information you provide can be edited and completed together with FIN-CLARIN.
- Completing the form does not oblige you to conclude the deposition agreement, but the information may be of great help if you need furthe advice on your resource later.
- FIN-CLARIN may also contact you, or the responsible person you have indicated, to agree on follow-up measures concerning the resource.
- Once you have requested that we add the metadata of the language resource in the language resource catalogue, your resource can immediately gain more visibility, even if it is not yet ready for publication.
- FIN-CLARIN is happy to help you with any questions related to the deposition and distribution of the resource. You can reach us by sending email to fin-clarin (ATT) helsinki.fi

Other language resources to be published by Kielipankki (The Language Bank of Finland) of FIN-CLARIN

Contact details

The e-form for describing a new resource for the Language Bank of Finland

(*) Name of the information provider *

(*) Email address of the information provider *

ORCID identifier of the information provider (instructions)

What is ORCID?

The home organization of the information provider, and the institution or department when appropriate; preferably also

Metadata


META SHARE
Elias Lönnrot Letters Online, source
View resource name in all available languages
lonnrot-src
Persistent Identifier of this resource: http://urn.fi/urn:nbn:fi:lb-2022040501
Access location: http://urn.fi/urn:nbn:fi:lb-2022040502
The resource is available in Kielipankki's download service www.kielipankki.fi/download
The corpus consists of the correspondence of Elias Lönnrot with private individuals as well as institutions
Elias Lönnrot was the creator of the Kalevala, medical doctor and... Read More
« Back
Edit Resource
Citable PID
Location PID
License
Distribution
Availability
Available - Unrestricted Use
Licence
CC - BY
Distribution Access/Medium:
Downloadable
Licensors:
The Finnish Literat
Distribution rights
The Finnish Literat
IPR Holder
The Finnish Literat
Contact Person
User support FIN-C
Documentation
Resource group page: http://urn.fi/urn:nb...
Document Type: Other
Lisenssi (Elias Lönnrotin kirjeenvaihto),
License (Elias Lönnrot Letters Online),
http://urn.fi/urn:nb...
Editor: FIN-CLARIN
how to cite: https://www.kielipan...

Clear the License and Acquire Source Data

- Negotiate on the license for distributing the resource
- Check for copyrighted content
- Check for personal data
- Deposition license agreement

- Receive the source data, check and describe

Clear the License and Acquire Source Data

- → Roles: Data depositor, Rightholder, Controller (for personal data)
- → Deposition license agreement template
- → End-user license template (+ license page template)
- → Template for resource-specific data protection terms and conditions
- → Language Bank Rights (LBR), to support CLARIN RES

CLARIN License Categories

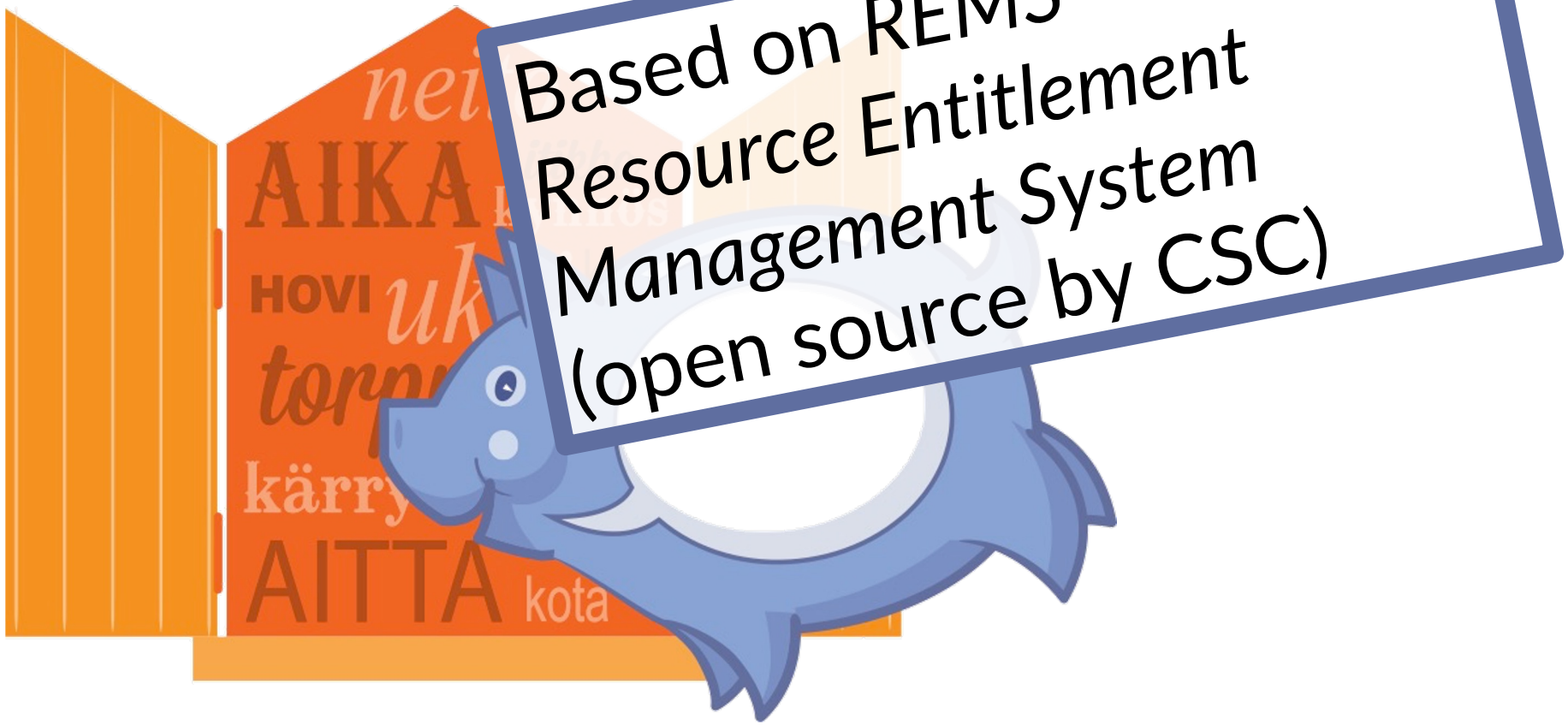
Publicly available

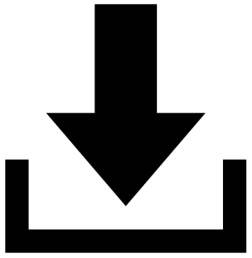
Available for academic, logged in users

Personal permission is required for access
Language Bank Rights, *https://lbr.csc.fi*

Language Bank Rights

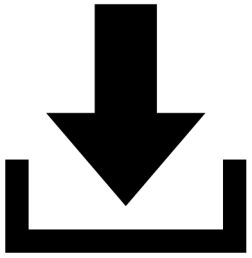
https://lbr.csc.fi





Publish the Source Data in Download

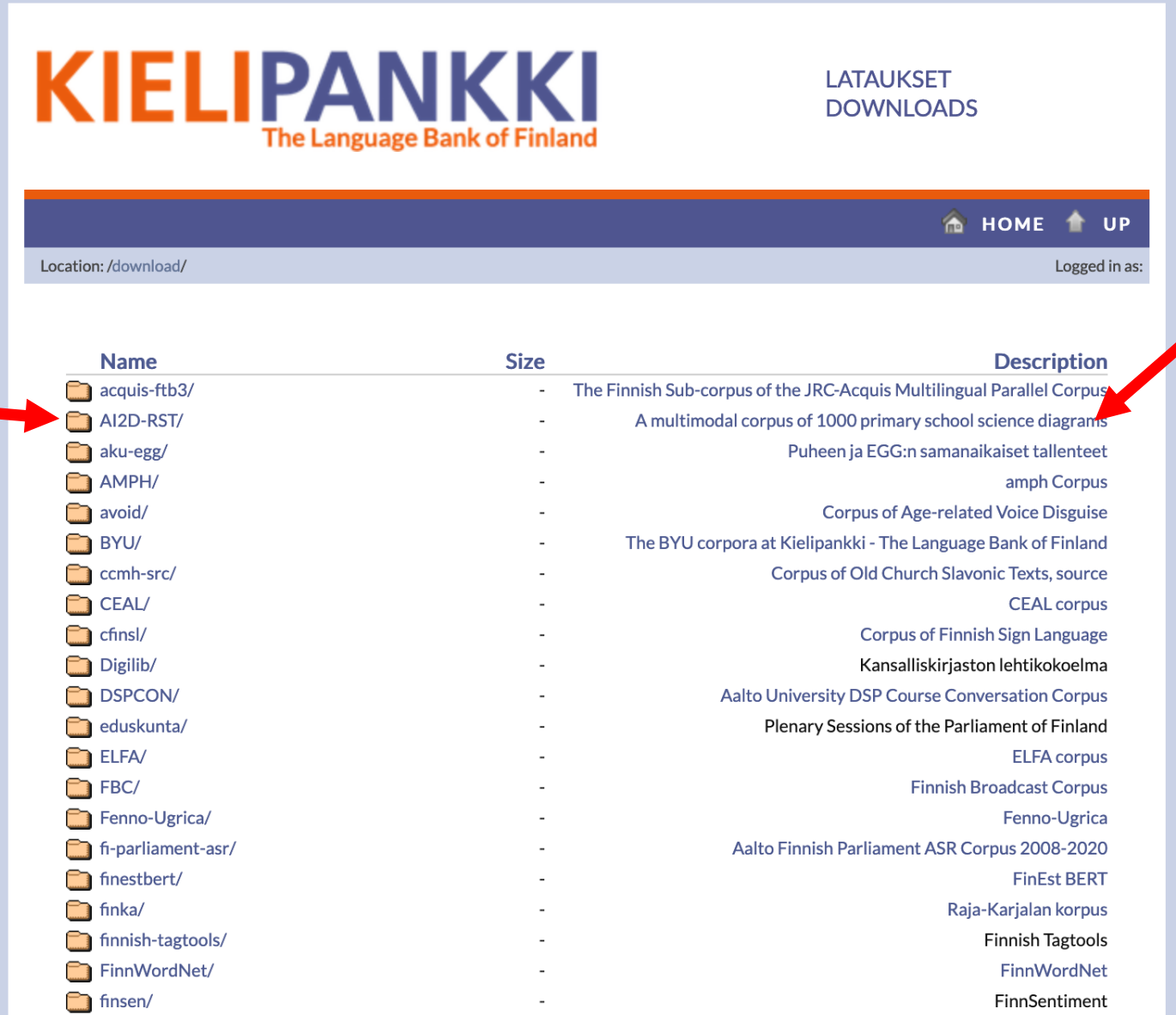
- Request a location PID
- Create README and LICENSE information files
- Package the data
- Create an LBR record (for RES licensed corpora)
- Upload the zip package(s) to the download service
- Publish a news item online



Publish the Source Data in Download

→ PID generator
→ Language Bank Rights (LBR)
→ Access rights to the download server
→ Editor rights in the Portal

Download Service, *http://www.kielipankki.fi/download*



Folder containing zipped data packages

Link to metadata

Publish the Data in Korp

- Preserve the structural features of the source data
- When possible, represent resource-specific additional properties as search criteria / filters
- Support links to media and/or external resources

KORP v9

25 of 275 corpora selected — 199.27M of 3.46G tokens

KIELIPANKKI The Language Bank of Finland

language..nn.1

Resource: *e-thesis*

Simple search: *language (noun)*

Simple | Extended | Advanced | Compare

language (noun) Search

in order and also as initial part compound_middle final part and case-insensitive

KWIC: hits per page: 25 sort within corpora: not sorted Statistics: compile based on: word Show statistics

KWIC | Statistics

Results: 58,667

« ‹ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ... › » Go to page of 2347 Show context

E-THESIS: DOCTORAL DISSERTATIONS: AGRICULTURE AND FORESTRY (does not support extended context)

alo Natural Resources Institute Finland (LUKE) FI-31600 Jokioinen, Finland	**Language**	revision: B.Sc. Michael Bailey Maininkitie 9 FI-02320 Espoo, Finland Cover: Kati Lassi, Mikko Hautala, Mikko Hakojärvi
Johannes Tiusanen has given to the content and	**language**	of the articles associated with this thesis.
I thank B.Sc. Michael Bailey for	**language**	revision of my thesis.
The research assistant spoke two of the local	**languages**	including Mòoré, which is spoken in approximately 70 of the village communities.
A research assistant who understood and spoke the local	**language**	led the discussion on tree planting in the region.
esses interpersonal skills such as the ability to communicate in the three local	**languages**	used in the communities.
Helsinki, Finland	**Language**	revision:
Stephen Stalter	**Language**	Services PO Box
Emotion inferences from vocal expression correlate across	**languages**	and cultures.
ls fermented milk products allowed the word yoghourt to enter in the common	**language**	(Baglio, 2014).
rgi, Eeva Blomqvist - Vijendran and Peter Seenan are thanked for editing the	**language**	of publications and this summary.
Sakari Alasuutari	**Language**	revision:
Fourth, the entire data gathering was implemented and conducted in local	**languages**	, which were the native languages of the interviewers, collaboration partners and company managers, but were translated into English afterwa
g was implemented and conducted in local languages, which were the native	**languages**	of the interviewers, collaboration partners and company managers, but were translated into English afterwards to avoid possible problems cau
were translated into English afterwards to avoid possible problems caused by	**language**	barriers.
I warmly thank MMT James Thompson for careful	**language**	editing, Birgitta Åhman for accepting the duty of opponent and Pekka Uimari, my custos, for support with all the administrative issues.
A	**language**	and environment for statistical computing.
menting on this Thesis, and also Stephen Skate for his revision of the English	**language**	.
As often when doing sensory research, translations between	**languages**	may be challenging (Andani et al.
ISSN 2242 - 119X (Print) ISSN 2242 - 1203 (Online) Date September 2015	**Language**	English Pages 101 p. + app. 41 p.
special thanks also to Stella Thompson, University of Helsinki, for revising the	**language**	of this dissertation.
Economics of Forestry (1902) - apparently the first on the subject in English	**language**	- that he was fully aware of the German tradition in forest economics.
nd loss, business partnership, and even of the taking of interest in their native	**language**	(Poitras 2000).
mmercial arithmetic tables ₶T⅋ became common also in other countries and	**languages**	.
The first compound interest tables in the English	**language**	were by William Colson (1612) whose book contained tables of ₸ꝺ at ten percent per annum, for ₶

CORPUS

E-thesis: Doctoral dissertations: Agriculture and Forestry

Subcorpus of: E-thesis (English)

Metadata

Cite corpus

Link to corpus in Korp

Persistent identifier: urn:nbn:fi:lb-2020031301

Licence: CC BY (CLARIN PUB)

TEXT ATTRIBUTES

text_title: Challenges in real-time precision farming: a case study of modelling biomass accumulation

keywords: agroteknologia

faculty: University of Helsinki, Faculty of Agriculture and Forestry, Department of Agricultural Sciences, Agrotechnology

subject: *[empty]*

thesis type: Doctoral dissertation (article-based)

WORD ATTRIBUTES

base form: language

baseform (compound boundaries): language

part of speech: noun

msd: Number=Sing

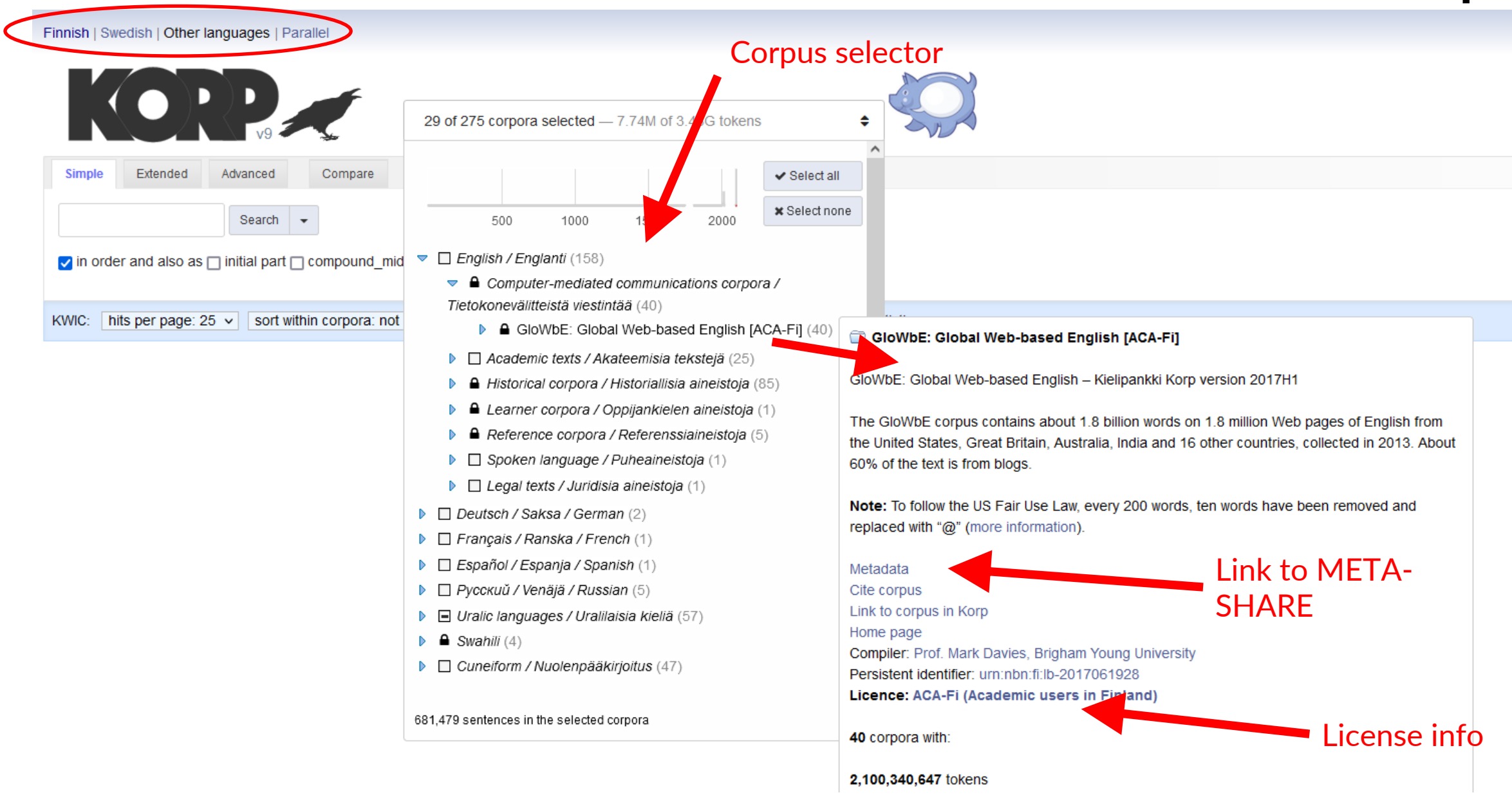
dependency relation: compound

Show Dependency Tree

Download from E-thesis

Resource metadata in Korp

Resources grouped after languages



www.kielipankki.fi

97._torstaina_23._lokakuuta_2008.mp4

Alkuaika: **10:00** Loppuaika: **10:12**

Puhunnos: **Sitten äskettäin Espoossa pidetyssä vaalitilaisuudessa ministeri Väyrynen nostatti pelkotiloja espoolaisissa todeten että metro tuo kaupunkiin maahanmuuttajia .**

eduskuntaryhmä: Sosialidemokraattinen eduskuntaryhmä
puhujan rooli: *[tyhjä]*
puheenvuoron tyyppi: *[tyhjä]*
puhuja: Maria Guzenina-Richardson /sd
puhunnoksen numero: 17661
puhunnoksen alkuaika (ms): 600928
puhunnoksen kesto (ms): 11357
puhunnoksen loppuaika (ms): 612285
istunnon kesto: 2:49:38,760
puhunnoksen alkuaika: 0:10:00,928
puhunnoksen loppuaika: 0:10:12,285
puhunnoksen kesto: 0:00:11,357

SANAN PIIRTEET

perusmuoto: maahanmuuttaja
perusmuoto (yhdyssanarajat): maahanmuuttaja
sanaluokka: substantiivi
morfologinen analyysi: NUM_Pll CASE_Par
dependenssisuhde: suora objekti
sanan sijainti nimessä: ulkopuolella (O)

Näytä dependenssipuu
Näytä video Korpissa
Näytä video erillisellä sivulla

Publish the Data in Korp

- Converting source data to simple XML (HRT)
- Tokenizing the data
- Parsing the data
- Inserting annotations
- Creating the corpus package for Korp
- Creating the corpus configuration for Korp
- Publishing a news item in Korp and in the Portal

Simple XML (HRT)

```
<text title="Elias Lönnrot &gt; Frans Johan Rabbe" author="Elias Lönnrot" lang="swe" datefrom="18510303" dateto="18510303">
<paragraph>Kajaani
</paragraph>
<paragraph>3.3.1851
</paragraph>
<paragraph>Käre Broder!
</paragraph>
<paragraph>Tack för katetrarne, som jag med posten erhöll. Med Öhmanska honorarii anbudet är jag mycket belåten och skall vid tillf
insända andra uppsattser. Affären om qvarliggande boklagret i öhmanska bokhandeln ber jag Dig afgöra efter eget godtfinnande
och de uppgifter, Du af nämnde Bokhandel erhåller, ty jag litar på att de äro riktiga och är nöjd med hvad som erbjudes.
Mycket smärtande för mig var Fab. Collans dödsfall, ty efter min
</paragraph>
<paragraph>öfvertygelse var han en i alla afseenden braf karl, och det finns ondt om sådana.
</paragraph>
<paragraph>Jag hade beslutit att med våra prester fara idag till Idensalmi för att gratulera nya Bispen, men en Ingenieur Kyanders
som ligger på sistone i Hydrocephalus acutus hindrar mig att resa.
</paragraph>
<paragraph>Dessa rader afsänder jag med min brorsson Frans, som nu i rappet afreser till Hfors för att studera, och hvilken jag på
bästa får hos Dig rekommendera.  Måtte han vandra varskoddare, än många andra i dessa kritiska tider!
</paragraph>
<paragraph>Kajana den
</paragraph>
<paragraph>3 Mars 1851
</paragraph>
<paragraph>Tuus
</paragraph>
<paragraph>Elias Lönnrot.
</paragraph>
```

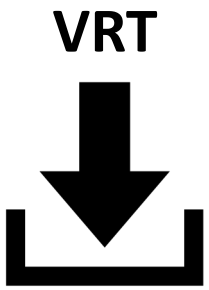
structural attributes

plain text inside paragraph tags

Publish the Data in Korp

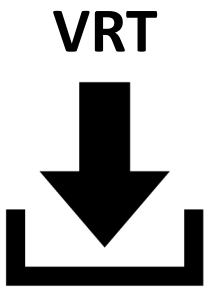
- → Tailor-made scripts regarding format and structure of the source data
- → Common tokenizer
- → Common parser
- → Additional tools, e.g., language identifier
- → Common script ’*korp-make*’
- → Manual additions to the existing Korp configuration

- → Access rights to the Korp server (within the team)
- → Access rights to the portal



Publish the VRT Data in Download

- Extract VRT from Korp
- Request location PID
- Create README and LICENSE information files
- Package the data
- Create an LBR record (for a RES corpus)
- Upload the zip package(s) to the download service
- Publish a news item



Publish the VRT Data in Download

→ Access rights to the Korp server
→ Tool for extracting VRT from Korp
→ PID generator
→ Language Bank Rights (LBR)
→ Access rights to the download server
→ Editor rights in the portal

VRT format, extracted from Korp

positional attributes

structural attributes

structural elements

```
<!-- #vrt positional-attributes: word ref lemma/ pos msd dephead deprel wid spaces lex/ -->
<!-- #vrt info: VRT generated from CWB data for corpus "nlfcl_sv" (2022-03-02 14:25:06 +0200) -->
<!-- #vrt info: A processing log at the end of file -->
<text id="1" author="" contributor="Hertzberg, Rafael" title="Gustaf Adolfs minne i Finland" year="1894"
lang="swe" datefrom="18940101" dateto="18941231" timefrom="000000" timeto="235959" natlibfi="Klassikkokirjasto,
rights="" digitized="2017-04-20" filename="KlK_with_timestamps.xml" urn="URN:NBN:fi-fd2010-00000511" pdflink="|1
html_url="http://www.doria.fi/handle/10024/100731" publisher="Helsingfors: [s.n.], 1894" license="Public Domain"
<page n="1">
<paragraph id="1">
<sentence id="1">
GUSTAF  1       |GUSTAF|        PM      PM.NOM  0       ROOT    1       _               |GUSTAF..pm.1|
II      2       |II|    RG      RG.NOM  1       ET      2       _       |II..rg.1|
ADOLF   3       |ADOLF| PM      PM.NOM  2       HD      3       SpaceAfter=No   |ADOLF..pm.1|
.       4       |.|     MAD     MAD     1       IP      4       SpacesAfter=\s\s\s      |...xx.1|
</sentence>
<sentence id="2">
Efter   1       |Efter| PP      PP      0       ROOT    1       SpacesAfter=\s\s\s      |Efter..pp.1|
ett     2       |den|en|        DT      DT.NEU.SIN.IND  3       DT      2       _       |den..al.1|en..al.1|
kopparstick     3       |kopparstick|   NN      NN.NEU.SIN.IND.NOM      1       PA      3       SpacesAfter=\s\s
af      4       |af|    PP      PP      3       ET      4       _       |af..pp.1|
M.      5       |M.|    PM      PM.NOM  4       PA      5       _       |M...pm.1|
J.      6       |J.|    PM      PM.NOM  5       HD      6       _       |J...pm.1|
Miereveld       7       |Miereveld|     PM      PM.NOM  5       HD      7       SpaceAfter=No   |Miereveld..pm.1
.       8       |.|     MAD     MAD     1       IP      8       SpacesAfter=\n\n        |...xx.1|
```

Check List of Tasks

corpus short name

ylenews-fi-2019-2021: Publish VRT data in Download

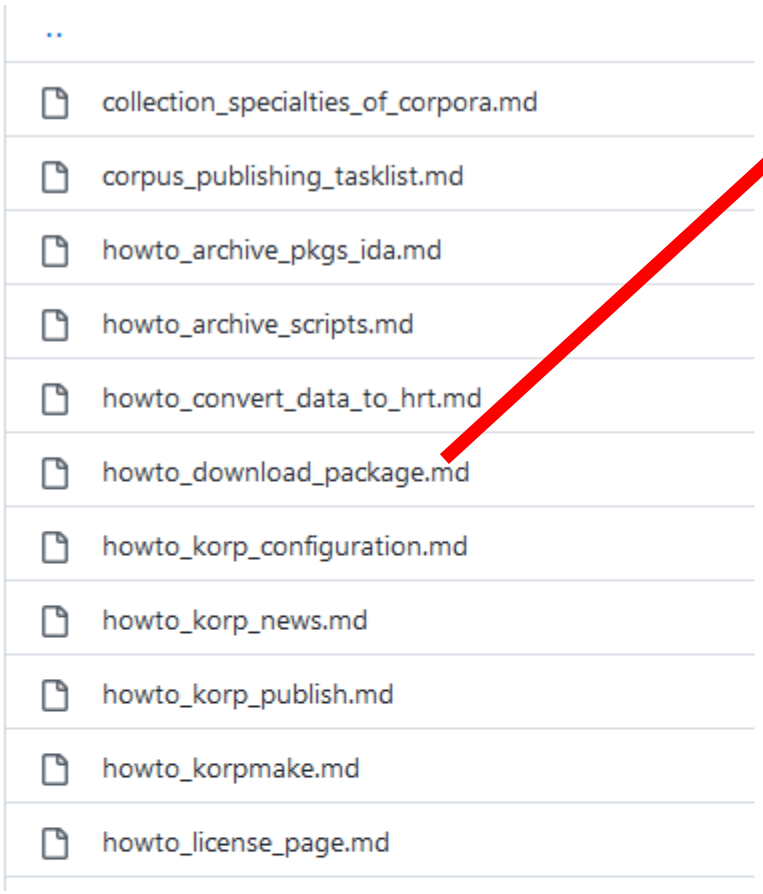
links to documentation modules

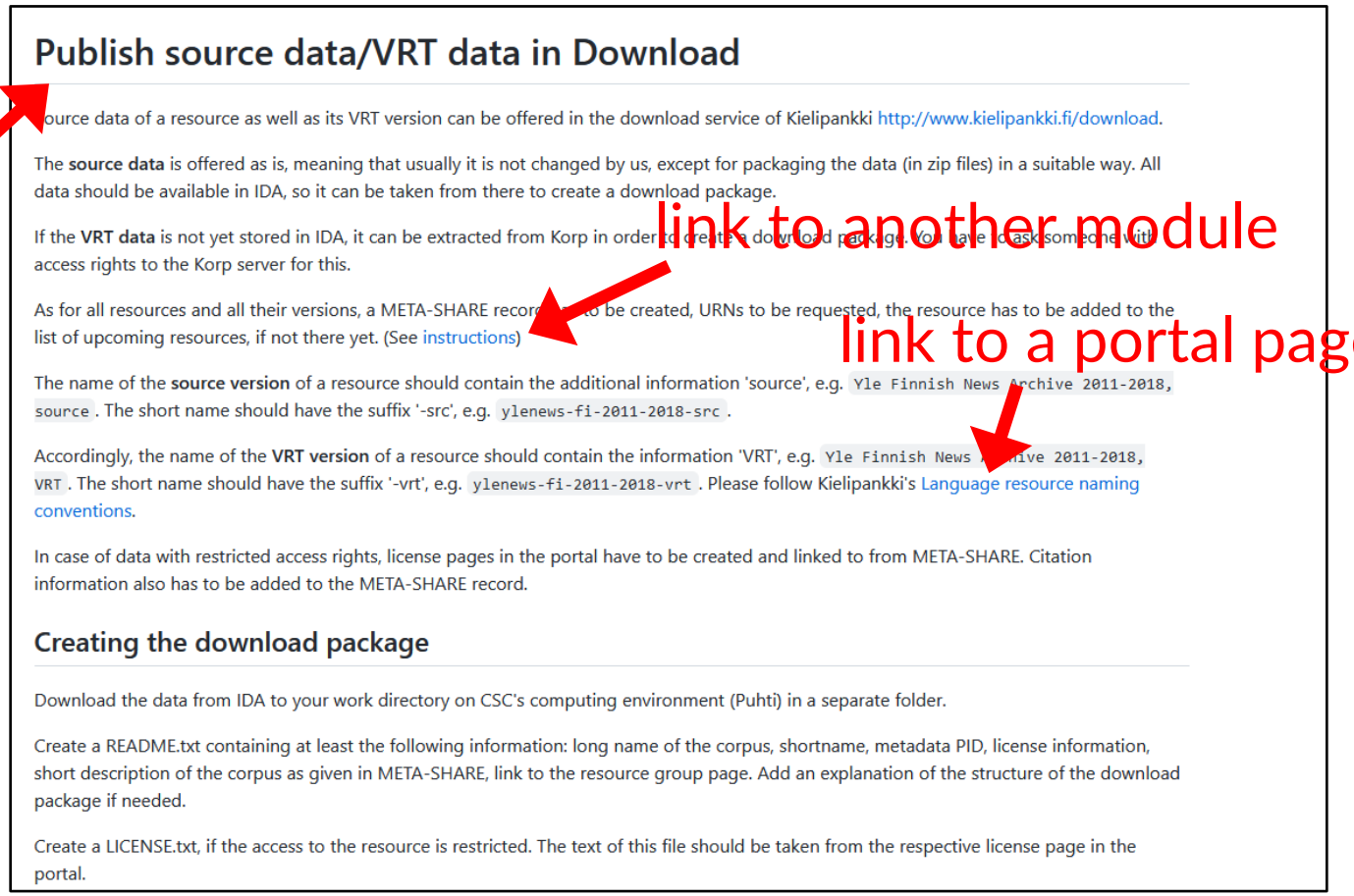
check boxes

1. [x] +***D*** Get the data from IDA or extract it from Korp
2. [x] +***A*** Create a META-SHARE record instructions for creating metadata records
3. [x] +***A*** Request URNs (for META-SHARE, download, license pages)
4. [x] +***A*** Add the corpus to list of upcoming resources
5. [x] +***A*** Create/update license pages how to create/update license pages
6. [x] +***A*** Add citation information to the META-SHARE record
7. [x] +***D*** Create a download package
    1. [x] +***D*** Create and add the downloadable readme and license files how to create/update license pages
    2. [x] +***D*** Zip the data and the readme and license files
    3. [x] +***D*** Compute MD5 checksum for the zip package
8. [x] +***D*** Add the download package, MD5 checksum file and readme and license files to the directory `/scratch/clarin/download_preview` on Puhti
9. [ ] +***T*** Have the package tested
10. [ ] +***D*** Upload the package to the download service (or ask someone with the rights to do that)
11. [-] ***?A*** Create an LBR record (for a RES corpus, if the corpus does not yet have one)
12. [ ] +***T*** Have it tested again (access rights!)
13. [ ] +***A*** Move the corpus from the list of upcoming resources to the list of published resources
14. [ ] +***A*** Update the META-SHARE record; add location PIDand Availability start date (under Distribution)
15. [ ] ***?A*** Update the META-SHARE record: add relations to previous or parallel versions/variants of the corpus
16. [ ] +***A*** Create or update the resource group page

Modular and Shared Documentation

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





link to another module

link to a portal page

Testing: Quality control

- the final part of the pipeline
- another team member checks the resource before it is published

https://www.kielipankki.fi/corpora

Etsi:

Abbreviation	Name and metadata	License	Apply	Location	Service level	Help	Cite
acquis-ftb3	The Finnish Sub-corpus of the JRC-Acquis Multilingual Parallel Corpus	PUB		Korp	B		
acquis-ftb3-dl	Finnish Sub-corpus of the JRC-Acquis Multilingual Parallel Corpus, Downloadable Version	PUB		Download	B		
agricola-v1-1-korp	The Morpho-Syntactic Database of Mikael Agricola's Works version 1.1, Korp	PUB		Korp	B		
ai2d-rst-v1-1	AI2D-RST: A multimodal corpus of 1000 primary school science diagrams version 1.1	PUB		Download	B		
aku-egg-dl	Speech and EGG (Electroglottography) Simultaneous Recordings, downloadable version	ACA		Download	B		
amph	amph-Corpus	ACA		Download	B		
ArkiSyn-korp	ArkiSyn Database of Finnish Conversational Discourse, Helsinki Korp Version	PUB		Korp	B		
AVOID	Corpus of Age-related Voice Disguise (AVOID)	RES		Download	B		
BeserCorp	The Corpus of Beserman Udmurt	PUB		Korp	B		
ccmh-src	Corpus Cyrillo-Methodianum Helsingiense: Corpus of Old Church Slavonic Texts, source	PUB		Download	B		
ceal-dl	The Downloadable Version of Classics of English and American Literature in Finnish	RES		Download	A		
ceal-o	Classics of English and American Literature in Finnish, Sentences and Paragraphs in the Original Order	RES		Korp	A		
ceal-par-korp	Classics of English and American Literature as translated by Kersti Juva, English-Finnish parallel corpus, Korp	RES		Korp	A		
ceal-par-s-dl	The Downloadable Version of Classics of English and American Literature as translated by Kersti Juva, English-Finnish parallel	ACA		Download			

link to metadata

link to license

link to location

link to resource group page

citation instructions

Resource group page

Wanca 2016

Wanca 2016 is a collection of web corpora in small Uralic languages. The collection is composed of 29 sentence corpora in different languages. The corpora have been collected from the Internet using the automated system developed in the Finno-Ugric Languages and the Internet project (SUKI) supported by the Kone foundation from their Language Programme 2012-2016. The sentences have been extracted from the pages found while harvesting with Heritrix and the language of each sentence has been identified with MultiLi using HeLI as the identification method. Each sentence has a link to the original page it was found in, but it is possible that some of the links stop working. In that case we recommend searching for the page in the Internet Archive Wayback machine https://archive.org/web/.

More information on Wanca: http://www.suki.ling.helsinki.fi/wanca

Latest versions/subcorpora:

Wanca 2016, Korp Version — Select the corpus in Korp
Metadata and license
Attribution instructions

Wanca 2016, source — Download the resource
Metadata and license
Attribution instructions

Wanca 2016, VRT — Download the resource
Metadata and license
Attribution instructions

Multiple versions of the same resource published in different means of publication

Search for these versions in META-SHARE

Of this language corpus different versions/subcorpora are published in the Language Bank of Finland. The versions are available through the Language Bank Download Service and/or through the Korp concordance tool. The links to the different versions can be found from the list above.

Kielipankki Resource Families

link to the resource group page of a resource or tool

Corpora

- Computer-mediated communication corpora
  - Corpus of Global Web-Based English
  - Finnish conversational chat corpus
  - FinnSentiment
  - SFNET Corpus
  - Suomi 24
  - TallVocabL2Fi: Measurements of 15 L2 Finnish le
  - The HS.fi News and Comments Corpus
  - Ylilauta Corpus
- Corpora of academic texts
  - UH's E-thesis corpus
- Easy-to-read corpora (or corpora containing easy-to-re
  - Corpus of Finnish Magazines and Newspapers fro
  - Yle News Archive
- Historical corpora
  - Alexis Kivi Corpus (SKS) [multilingual corpus]
  - Classics Library of the National Library of Finland
  - Classics of Finnish Literature, Kielipankki version

Tools

- Language identificatio
  - HeLI-OTS
- Named entity recognition
  - Finnish Tagtools
- Machine learning
  - FinBERT
- Part-of-speech tagging and le
  - Finnish Tagtools
  - Turku-neural-parser pip
  - Sparv
- Tools for sentiment analysis
  - Etuma Customer Feedb
  - finsentiment
- Tools for speech analysis and a
  - Aalto University Automa
  - ELAN
  - Praat

Web Services

- Data Set Location and Access
  - Digital collections from the National Library of Finland
  - Digital collections of Kotus, the Institute of the Languages of Finland
  - Download service Kielipankki
  - Giellatekno
  - Language Bank Rights Kielipankki
  - META-SHARE Kielipankki
  - OPUS Helsinki
  - Virtual Language Observatory CLARIN
- Interactive Content Search and Visualization
  - CLARIN Federated Content Search
  - Korp
  - Lääketutka
  - Texthammer
  - Text reuse in the Swedish-language press, 1645-1918
- Interactive Lexical Platforms
  - ANEE lexical portal of Akkadian
  - Proto-Indo-European Lexicon
  - Sanat
  - Signbank

Citation instructions

amph	amph-Corpus	ACA		Download	B		
ArkiSyn-korp	ArkiSyn Database of Finnish Conversational Discourse, Helsinki Korp Version	PUB		Korp	B		
AVOID	Corpus of Age-related Voice Disguise (AVOID)	RES		Download	B		
BeserCorp	The Corpus of Beserman Udmurt	PUB		Korp			
ccmh-src	Corpus Cyril… Slavonic Text…						

[suomeksi] [in English]

Reference instructions: AVOID

Please cite the language resource as follows:

Kinnunen, T., Hautamäki, R. G., Sahidullah, M., Hautamäki, V., Werner, S., & Bentz, M.. *Corpus of Age-related Voice Disguise (AVOID)* [speech corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2018060621

Show: [Bibtex] [Zotero]

Search for references to the language resource in Google Scholar

https://www.kielipankki.fi/news/

News

New resource: Open Richly Annotated Cuneiform Corpus, Korp Version, June 2021

12.9.2022

New resource: Open Richly Annotated Cuneiform Corpus, Korp Version, June 2021 Oracc 2021 (Open Richly Annotated Cuneiform Corpus, Korp Version, June 2021) is now available in Korp alongside Oracc 2019. [...]

New resource: Finnish conversational chat corpus, source

6.9.2022

New resource: Finnish conversational chat corpus, source Finnish conversational chat corpus, source is available at the download service at Kielipankki. More information can be found on the resource group page.

Researcher of the Month: Mikko Laitinen

5.9.2022

Researcher of the Month: Mikko Laitinen Photo: Olli Laitinen Kielipankki – The Language Bank of Finland is a service for researchers using language resources. Mikko Laitinen tells us about his [...]

Search the Language B
Haku ...
Subscribe to the ne
Kielipankki
Seuraa sivu
Kielipankki
Bank of Fin
noin viikko sitten
Kielipankin kuukaude
Laitinen, joka kertoo
viimeaikaisesta työst
parissa. Lyhyet tviitit
rikkaaseen metadata
uudenlaisen ikkunan
https://www.kielipankk
tutkija-mikko-laiti.../
#UEF #korpus #FINC
#tutkimus #aineisto #
#digihum

Active Role in Publishing a Resource?

PROS:

- A wide range of resources can be offered (including RES and even very large data sets)
- Clear licenses – clear responsibilities
- Consistent metadata
- Support for individual features in source data
- Consistent, transparent and interoperable data

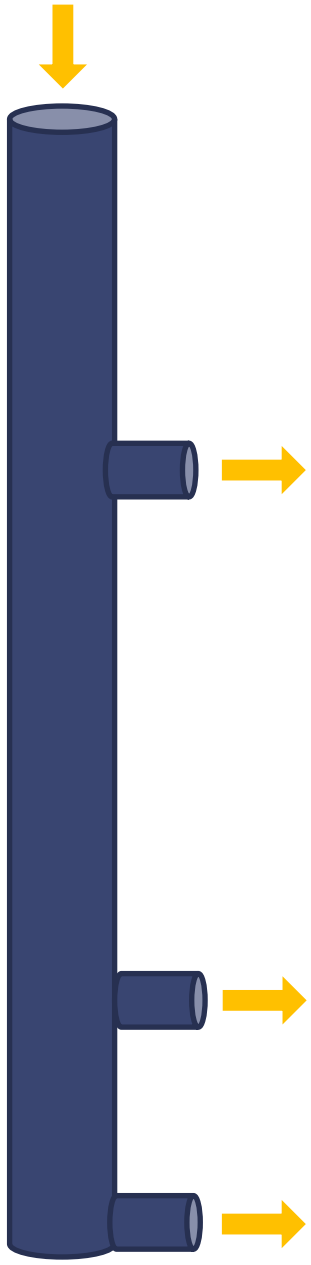
CONS:

- Workload
- Publication delay

Working days required in total

≥ 1

≤ 60 (?)



Enter a new resource

Clear the license

negotiate a deposition agreement

acquire the source data

Publish the source data in Download

convert to HRT

tokenize and parse

package the corpus for Korp

configure the corpus for Korp

Publish the annotated data in Korp

Extract VRT from Korp

Publish the VRT data in Download

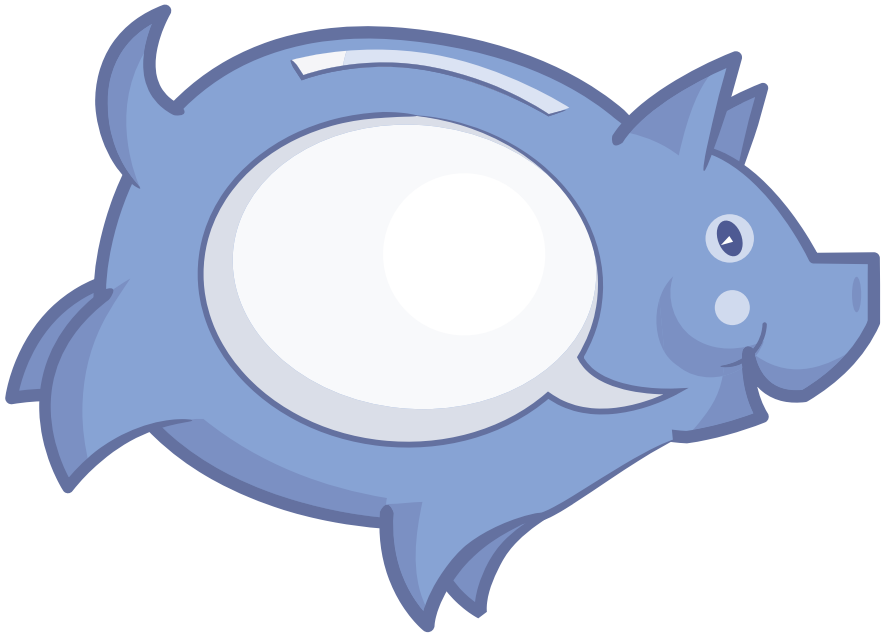
KIELIPANKKI
The Language Bank of Finland

Conclusions

- The shared check list helps in monitoring the publishing process.
- The modular documentation and shared scripts ensure consistent processing within the team.
- Increased automatization and publicly available documentation can enable the content depositors to participate more actively.
- It is necessary to compare and to share good practices with other CLARIN centres.

FIN-CLARIN

KIELIPANKKI
The Language Bank of Finland



Thank you!