KIELIPANKKI
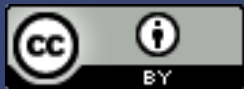## The Language Bank of Finland

Licensing and distributing language resources via the Language Bank of Finland

Mietta Lennes

FIN-CLARIN

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			🙿

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... ...ethodi... ...Helsi... ...Corpus of Old Sla... 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

Resources may have several versions or variants, potentially under different licenses!

Suomi24-2001-2015	The Suomi 24 2001-2015 (Sample) Corpus	ACA	Download	A	❷	🟇
suomi24-2001-2017-korp-v1-2	The Suomi24 Sentences Corpus 2001-2017, Korp version 1.2	PUB	Korp	A	❷	🟇
suomi24-2001-2017-vrt-v1-1	The Suomi24 Corpus 2001-2017, VRT version 1.1	ACA	Download Puhti	A	❷	🟇
suomi24-2001-2020-korp	The Suomi24 Sentences Corpus 2001-2020, Korp version	PUB	Korp	A	❷	🟇
suomi24-2001-2020-vrt	The Suomi 24 Corpus 2001-2020, VRT version	ACA	Download	A	❷	🟇
Suomi24-2015H1	The Suomi 24 Corpus (2015H1)	ACA	Download	A	❷	🟇
Suomi24-2016H2	The Suomi 24 Corpus (2016H2)	ACA	Download	A	❷	🟇
suomi24-2018-2020-korp	The Suomi24 Sentences Corpus 2018-2020, Korp version	PUB	Korp	A	❷	🟇
suomi24-2018-2020-vrt	The Suomi24 Corpus 2018-2020, VRT version	ACA	Download	A	❷	🟇

Resource-specific metadata

META=SHARE

The Suomi 24 Sentences Corpus 2001-2020, Korp version

⊙97 ✓4

▸ View resource name in all available languages

suomi24-2001-2020-korp

Persistent Identifier of this resource: http://urn.fi/urn:nbn:fi:lb-2021101525

Access location: http://urn.fi/urn:nbn:fi:lb-2021101526

This content is available in Kielipankki.

This collection contains two sets of Suomi24 data: "The Suomi24 Sentences Corpus 2001-2017, Korp version" and "The Suomi24 Sentences Corpus 2018-2020, Korp version".
Together, the two corpora cover all the discussion forums of the Suomi24 online social... Read More

« Back   Edit Resource

Access location

text 📄

Distribution
Availability
Available - Restricted Use
Licence
CC - BY - NC
Restrictions: Academic - Non Commercial Use, Attribution
User Nature: Academic
Distribution Access/Medium: Accessible Through Interface

Monolingual text corpus
Languages
Finnish
Linguality
Linguality type: Monolingual
Size
1
Modalities
Written Language

Metadata
Created: 10/15/2021
Last Updated: 01/31/2022
Revision: location PID added
Metadata Creator
Ute Dieckmann 📧
Mietta Lennes 📧
Relation
Related Resource: The Suomi24 Sentences Corpus 2018-2020, Korp version,

yhdessä valtion ja seutujen kuntien kanssa pilottiohjelma **maahanmuuttajien** kotouttamisen ja työllistymisen edistämiseksi.

Samoin on koko yhteiskunnan etu myös että **maahanmuuttajat** työllistyvät mahdollisimman hyvin ja saavat tarvitsemansa

Sinne tuli sekä paikallisia että näitä **maahanmuuttajia** pienen ajan päästä tuli valtavasti kumpienkin vanhempia

keskeisessä asemassa ja tiedämme

ksi olisivat kuin vuokratyöyrityksiä

uomen kielen opetusta on tarjolla ja

n aikaan meillä sanotaan että maa

pelkotiloja espoolaisissa todeten e

kunta teki yksitoista toimenpide eh

ssa eduskunta edellytti hallitusta ry

1 2 3 4 5 6 7 8 9 10

JSON

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

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 .**

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

KIELIPANKKI
The Language Bank of Finland

LATAUKSET
DOWNLOADS

HOME   UP

Location: /download/

Logged in as:

Name	Size	Description
acquis-ftb3/	-	The Finnish Sub-corpus of the JRC-Acquis Multilingual Parallel Corpus
AI2D-RST/	-	A multimodal corpus of 1000 primary school science diagrams
aku-egg/	-	Puheen ja EGG:n samanaikaiset tallenteet
AMPH/	-	amph Corpus
avoid/	-	Corpus of Age-related Voice Disguise
BYU/	-	The BYU corpora at Kielipankki - The Language Bank of Finland
ccmh-src/	-	Corpus of Old Church Slavonic Texts, source
CEAL/	-	CEAL corpus
cfinsl/	-	Corpus of Finnish Sign Language
Digilib/	-	Kansalliskirjaston lehtikokoelma
DSPCON/	-	Aalto University DSP Course Conversation Corpus
eduskunta/	-	Plenary Sessions of the Parliament of Finland
ELFA/	-	ELFA corpus
FBC/	-	Finnish Broadcast Corpus
Fenno-Ugrica/	-	Fenno-Ugrica
fi-parliament-asr/	-	Aalto Finnish Parliament ASR Corpus 2008-2020
finestbert/	-	FinEst BERT
finka/	-	Raja-Karjalan korpus
finnish-tagtools/	-	Finnish Tagtools
FinnWordNet/	-	FinnWordNet
finsen/	-	FinnSentiment

Downloadable
speech corpora:

Local use in Praat

Downloadable speech corpora:

Local use in ELAN

CLARIN licence categories

PUB Publicly available

ACA Available for academic, logged in users

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

More detailed license conditions

+BY   author must be cited
+NC   non-commercial use only
+ID   login is required
+PLAN   research plan is required
+**PRIV** **contains personal data**
+NORED   redistribution is not allowed
+DEP   modified versions can be redistributed via CLARIN

and other resource-specific conditions, if required (e.g., data protection terms and conditions)

Why we need to develop our "pipeline"

- Over 260 individual resources currently available
- Over 100 forthcoming resources on the waiting list
  → need for managing and monitoring the publication process
- Resources vary in types of content: text, audio, video, lexical etc.
- Multiple services may require pre-processing steps: download, Korp, others (e.g., the Sensitive Data platform)
- License conditions vary, personal data may be included
- Division of labour between the depositor and Kielipankki staff: different knowledge, expertise and access privileges

Two types of agreements are usually needed:

1. Deposition (distribution) license agreement (DELA)
   - The Language Bank (University of Helsinki) obtains the right to distribute the resource to end-users under specific terms and conditions.
   - a DELA is usually required unless the University of Helsinki is the rightholder or the resource is already available under a public license.

2. End-User License Agreement (EULA)
   - The End-User agrees to use the data under specific conditions which the rightholder has approved.

CLARIN End-User License v2.1

- Updates to the end-user license templates (FI+ENG): https://www.kielipankki.fi/support/clarin-eula

- The most important additions and changes:
  - more consistent terminology in English and in Finnish
  - better coherence with the deposition license agreement template
  - possibility to include resource-specific data protection terms and conditions for resources containing personal data
  - right to modify the resource-specific license conditions after notifying users

CLARIN RES end-user license +PRIV +OTHER v2.1

Resource: Example Resource (URN: urn:nbn:fi:lb-1234597890)

Rightholder: Tiina Tutkija

The Rightholder grants the End-User a personal free, non-exclusive and perpetual (for the duration of the copyright) right to use and make copies of the Resource for the purpose agreed with the End-User as such, as modified, or as part of a compilation or derived work. The permission applies to all known or future modes and means of communication and includes a right to make modifications enabling the use of the Resource on other devices and in other formats.

Additional license terms as defined in the Terms of Service Agreement:

Identification and Access Conditions

- **ID**: The End-User must be authenticated or identified.
- **PLAN**: The End-User must present a research plan before the license can be granted. The End-User may only use the Resource for the purpose described in the research plan.

General conditions of use

- **BY**: The Author(s) of the Resource must be mentioned in connection with use.
- **PRIV**: There are personal data in the Resource. The End-User must comply with the data processing terms and conditions of this Resource. The End-User may process the personal data in the Resource only as long as there is a legal purpose for the processing, after which the personal data must be deleted.

Distribution conditions

- **NORED**: The Resource may not be disclosed to third parties. However, the Resource may be disclosed to subcontractors with whom appropriate agreements have been made concerning the subcontracted service and the processing of personal data. Subcontractors are only allowed to use the Resource to perform the service.

Other terms of use

- **OTHER**: There are other non-standard conditions in the license that the End-User should pay attention to.

  Prior to accepting this agreement and starting to process the Resource, the End-User must ensure that it has the right to accept the license and the data protection terms and conditions according to the instructions provided by the End-User's home organization.

  Data protection terms and conditions for this Resource: [link and/or persistent identifier of the resource-specific data protection terms and conditions]

This license has been made in compliance with copyright agreements by WIPO - the World Intellectual Property Organization. The rights granted in this license shall be so interpreted that in case applicable intellectual property laws grant rights not mentioned in this license, they are also regarded as part of the rights to be licensed; the purpose of this license is not to restrict any rights intended to be licensed within different legal systems. Additional rights to the Resource may be agreed separately in writing.

The Language Bank has the right to amend these terms on legitimate grounds, which may relate to, including but not limited to, instructions given by authorities, best practice, or changes in data protection laws or other applicable laws. A notification of any changes made shall be provided sixty (60) days before they take effect on the Language Bank's website, as well as by email to the addresses included in the User's application for access to the Resource.

Last updated : 24.5.2022

Example license

(this license includes resource-specific data protection terms and conditions)

Deposition license agreement template

- Updated deposition license agreement templates (Finnish, English): https://www.kielipankki.fi/support/dela/

- Significant additions and changes:
  – more consistent terminology in English and in Finnish
  – better coherence with the end-user licenses

Structure of the deposition license agreement

- Signature part; parties of the agreement:
  - Language Bank (= University of Helsinki)
  - Rightholder
  - Controller (the Data Controller of the personal data in the original resource)
- **Annex 1: General terms** (not to be modified for individual resources)
- **Annex 2: Identification of the Resource**, specifications and referencing convention
- **Annex 3: License** chosen for the agreement (CLARIN PUB / ACA / RES + specific conditions)
  - *Option 1*: The Rightholder and Controller submit the Resource to the Language Bank for permanent administration
  - *Option 2*: The license is valid only as long as the Agreement remains in force. The Rightholder (+ Controller if applicable) decides on the disclosure of the Resource
  - *Option 3*: The Rightholder wishes to make the Resource available under a public license (e.g., Creative Commons) and the content is to be made available for the first time via the Language Bank, e.g., for convenient discovery and maintenance.
- **Annex 4: Parts licensed by third parties**
- **Annex 5: Description of the Resource containing personal data and the data protection terms and conditions** applicable to the End-User
- **Annex 6: Data Protection Annex** (included only if selecting *Option 2*, where the Language Bank is to act as processor of personal data on behalf of the data controller)

The depositor's responsibilities

- Prepare the content of the resource so that the Language Bank and other researchers can understand it.

- Follow good practices in research ethics (ethical review?)

- **Follow the instructions given by your home organization.**

- Make sure you have the rights to distribute the data for research purposes

  - Copyrights: applicable license; or specific permission from third parties if needed

  - Personal data: data protection requirements should be cleared with your home university

    - Remember to inform the research subjects about potential re-use (**Kielipankki brochure**)

    - Assess the risks. Is DPIA (Data Protection Impact Assessment) required?

- Anonymize or pseudonymize the data if appropriate, considering further research use.

Contact FIN-CLARIN for details:
https://www.kielipankki.fi/support/   *fin-clarin@helsinki.fi*

Deposition workflow at the Language Bank 1/2

DEP = Depositor of the resource

CTRL = Data Controller (of the personal data)

ADMIN = Administrative staff in the Language Bank

UHEL = University of Helsinki (officially representing the Language Bank)

1. DEP: Submit preliminary metadata and background info (e-form)
2. ADMIN: Check the preliminary metadata and assess if the resource could be distributed; DEP: check metadata and confirm, ADMIN: obtain a persistent identifier as the anchor point for further discussion regarding the resource.
3. ADMIN: Ask depositor about the information given to the subjects, potential DPIA, and third-party permissions for sharing the resource; DEP+CTRL: produce the details and documents; ADMIN: suggest a license.

Deposition workflow at the Language Bank 2/2

DEP = Depositor of the resource
CTRL = Data Controller (of the personal data)
ADMIN = Administrative staff in the Language Bank
UHEL = University of Helsinki (officially representing the Language Bank)

4. ADMIN: Prepare a draft of the deposition license agreement and send to depositor; DEP: check details.
5. ADMIN+DEP: Have a meeting to finalize the agreement, if required.
6. DEP+CTRL: make decisions about licensing options and data protection conditions;
   DEP+ADMIN: complete the agreement text.
7. DEP (+CTRL) + UHEL: Sign the agreement.
8. DEP: Submit the data to the Language Bank for publication.

More about the resource publishing pipeline:

- Ute Dieckmann, Mietta Lennes, Jussi Piitulainen, Jyrki Niemi, Erik Axelson, Tommi Jauhiainen and Krister Lindén:
  The Resource Publishing Pipeline of the Language Bank of Finland.

  (CLARIN Annual Conference 2022, Day 3, Wednesday 9:20-9:40)

FIN-CLARIN
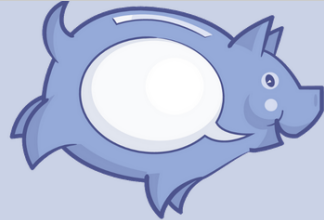
Language Bank Rights
https://lbr.csc.fi

Log in at *https://lbr.csc.fi*

KIELIPANKKI
The Language Bank of Finland

Login

Login with Haka credentials

Other login options

Select resources

EN FI   👤 Mietta Lennes / **Sign out**

Resource		
Aalto University DSP Course Conversation Corpus 2013-	More info 🔗	Add to cart
Amph Corpus	More info 🔗	Add to cart
BYU corpora (COCA, COHA, GloWbE), Downloadable Version	More info 🔗	Add to cart
Chuvash Corpus (UHLCS)	More info 🔗	Add to cart
Classics of English and American Literature in Finnish	More info 🔗	Add to cart
Classics of English and American Literature in Finnish, Scrambled Paragraphs	More info 🔗	Add to cart
Corpus of Age-related Voice Disguise	More info 🔗	Add to cart

7.11.2022

Fill in the application and accept license

Resources

Corpus of Age-related Voice Disguise — **More info** ↗

Change resources...

Terms of use

Each member is required to accept the terms of use to access the resources.

Kielipankki's General Terms of Use ↗

CLARIN RES +NC +PRIV +DEP (AVOID) ↗

Accept the terms of use

Application

Short Title of Research (max 80 chars) *

Abstract of Research Plan or Short Summary of Purpose of Use *

Actions

Save as draft   **Send application**   **Close application...**

Copy as a new application

In case the resource contains personal data, the license may also include further terms and conditions regarding the processing of personal data.

Publish the link to the Privacy Notice regarding the processing of personal data in your research

A general description can help researchers, research subjects and the larger public in understanding for which purposes corpora are used and how this research can benefit the society.

Notice on the processing of personal data included in a resource available in the Language Bank of Finland

Suomeksi

When using a resource that contains personal data and that is available in the Language Bank of Finland, you may use this form for submitting public information about the processing of the personal data. **NB. For the time being, this form can only be used for corpora that are currently available in the Language Bank of Finland.** Please do not submit details, e.g., about a forthcoming corpus that you are still collecting.

Public information about the purpose of use of the corpus or resource

The official title and persistent identifier can be found on the metadata page of the corpus or, e.g., on the corpus list of the Language Bank of Finland, https://www.kielipankki.fi/corpora.

Title of the corpus *?* *

Persistent identifier (URN) of the corpus *?* *

General title of the research topic or project, in Finnish: *?* *

General title of the research topic or project, in English: *?* *

Guidelines for processing corpora stored in the Language Bank of Finland that contain personal data

URN for this page: http://urn.fi/urn:nbn:fi:lb-2020081522

Always comply with these guidelines when processing corpora obtained from the Language Bank of Finland that contain personal data.

Does the corpus contain personal data?

Corpora stored in the Language Bank of Finland that contain personal data have the following label in their licence:

> **PRIV:** *There are personal data in the resource.*

The licence details of individual corpora can be found in the corpora listing of the Language Bank of Finland next to the corpus in question as well as in its metadata, which can be accessed using the persistent identifier assigned to the corpus (i.e., the URN address included in the citation instructions).

Resource-specific data protection terms and conditions

All corpora labelled PRIV contain a separate description of the **resource-specific data protection terms and conditions**, including the following details:

- Data controller of the personal data that are distributed via the Language Bank of Finland
- Types of personal data and data subject groups included in the corpus
- Description of the purposes for which the corpus can be further distributed by the Language Bank of Finland
- Restrictions regarding the location and transfer of the personal data to countries outside Finland
- Further processing instructions pertaining to personal data in the specific corpus, if any

General guidelines for processing personal data

Benefits of using a dedicated CLARIN centre

- persistent storage and referencing
- avoid keeping unnecessary copies "just in case"
- shared licensing, data security
- log: keeping track of who is using/accessing a given data set
- the metadata and documentation of related resources can be linked → a suitable variant is easy to locate (→ less need to process a "full version" with identifiers?)

KIELIPANKKI
The Language Bank of Finland
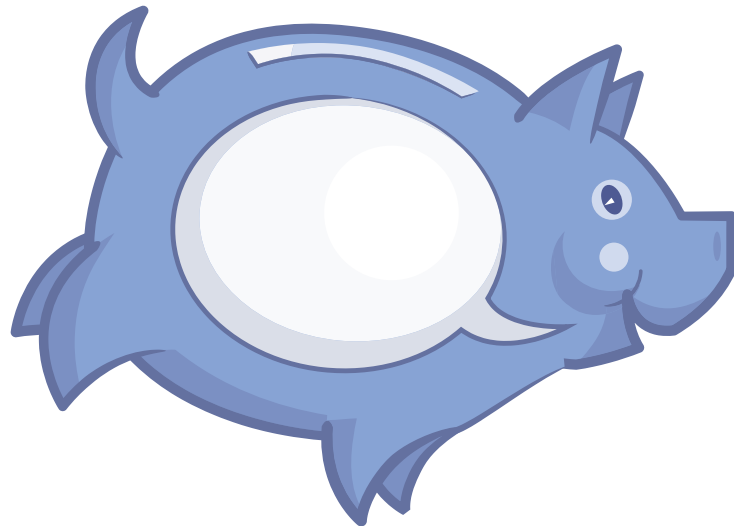
www.kielipankki.fi

General user support
fin-clarin@helsinki.fi

Technical services, user accounts etc.
kielipankki@csc.fi

Kiitos! Tack! Thank you!