

KIELIPANKKI
The Language Bank of Finland

Kielipankki pähkinäkuoreessa The Language Bank of Finland in a nutshell

CSC User Support Coffee 30.11.2022

Martin Matthiesen, CSC



At a glance

- Storage and sharing of language resources
- Tools for researchers and teaching
- 14+ billion tokens of text, 1000 hours of audio/video data.
- Licensing
- Access control (Open data done right)



www.kielipankki.fi

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.

What is the Language Bank of Finland?
The Language Bank of Finland is a service for researchers using language resources across digital humanities and social sciences. The Language Bank has a wide variety of text and speech corpora and tools for studying them. The corpora can be analysed and processed with the Language Bank's tools or downloaded.
Many corpora are publicly accessible, some require logging in. The rights to use restricted resources can be applied for electronically. Using the Language Bank is free for researchers and students.
The Language Bank is coordinated by the national FIN-CLARIN consortium formed by Finnish universities and other research organizations. FIN-CLARIN is a part of the international CLARIN ERIC research infrastructure. FIN-CLARIN enables access to the whole CLARIN community's language resources. Researchers and research groups can also agree with FIN-CLARIN upon depositing and distributing their own material.
If you are new to the Language Bank, take a look at the Language Bank introduction.

Who are the Language Bank's users?
The Language Bank's users are researchers, teachers and students who are interested in text, speech or video corpora containing natural languages and tools for utilizing them. Majority of the Language Bank's users are language researchers but the service is equally well suited to other digital humanities research.
For practical examples of the Language Bank's possibilities, see our monthly researcher portraits.

Search the Language Bank Portal:
[Search bar with 'has' entered]

Researcher of the Month: Marjatta Palander

News

- Are you planning to use speech or text materials in your MA or PhD thesis? Join the online course Data Clinic! (7.11.2022)
- Join the online course Introduction to Speech Analysis! (7.11.2022)
- Researcher of the Month: Marjatta Palander (6.11.2022)
- Donate Speech corpus available for research use - and soon for companies, too! (28.10.2022)
- New resource: Corpus of Contemporary American English - Kielipankki download version 2020 (28.10.2022)

More news

Contact
The Language Bank's technical support:
Kielipankki (ei) cc:0
tel: +358 9 4372001

Requests related to language resource:
Kielipankki (ei) cc:0
tel: +358 22 4122117

More contact information

CLARIN CENTRE B

FIN-CLARIN



Language Resources

www.kielipankki.fi/corpora

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	?	99
acquis-ftb3-dl	Finnish Sub-corpus of the JRC-Acquis Multilingual Parallel Corpus, Downloadable Version	PUB		Download	B	?	99
agricola-v1-1-korp	The Morpho-Syntactic Database of Mikael Agricola's Works version 1.1, Korp	PUB		Korp	B	?	99
ai2d-rst-v1-1	AI2D-RST: A multimodal corpus of 1000 primary school science diagrams version 1.1	PUB		Download	B	?	99
aku-egg-dl	Speech and EGG (Electroglottography) Simultaneous Recordings, downloadable version	ACA		Download	B	?	99
amph	amph-Corpus	ACA	+	Download	B	?	99
ArkiSyn-korp	ArkiSyn Database of Finnish Conversational Discourse, Helsinki Korp Version	PUB		Korp	B	?	99
AVOID	Corpus of Age-related Voice Disguise (AVOID)	RES	+	Download	B	?	99
BeserCorp	The Corpus of Beserman Udmurt	PUB		Korp	B	?	99
ccmh-src	Corpus Cyrillo-Methodianum Helsingiense: Corpus of Old Church Slavonic Texts, source	PUB		Download	B	?	99
ceal-dl	The Downloadable Version of Classics of English and American Literature in Finnish	RES		Download	A	?	99
ceal-o	Classics of English and American Literature in Finnish, Sentences and Paragraphs in the Original Order	RES	+	Korp	A	?	99
ceal-par-korp	Classics of English and American Literature as translated by Kersti Juva, English-Finnish parallel corpus, Korp	RES		Korp	A	?	99
ceal-par-s-dl	The Downloadable Version of Classics of English and American Literature as translated by Kersti Juva, English-Finnish parallel corpus, scrambled	ACA		Download		?	99
ceal-par-s-korp	Classics of English and American Literature as translated by Kersti Juva, English-Finnish parallel corpus, scrambled, Korp	ACA		Korp	A	?	99
ceal-s	Classics of English and American Literature in Finnish, Scrambled Paragraphs	ACA		Korp	A	?	99
cfnl-s-conv-dl	Corpus of Finnish Sign Language: conversations, Download version	RES	+	Download	B	?	99

metashare.csc.fi

META SHARE

Corpus of Age-related Voice Disguise

* View resource name in all available languages

AVOID

Persistent Identifier of this resource: <http://urn.fi/urn:nbn:fi:bn-201806062>

Access location: <http://urn.fi/urn:nbn:fi:bn-201901163>

This corpus includes normal and age-related disguised speech uttered by 60 native Finnish speakers (31 females and 29 males). The speakers were asked to read the same text fragments several times, in their modal voice and in two disguised voices, first proceeding to be an elderly speaker and then... [Read More](#)

* View resource description in all available languages

[Back](#) [Edit Resource](#)

Distribution

Availability
Available - Restricted Use

License

CLARIN RES
Restrictions: Academic - Non Commercial Use, Introduction, No Redistribution, Other
Attribution Details: Tomi Kinnunen, Rosa González Huartamäki, Mäe Sahoulian, Ville Huutaniemi, Stefan Wiener and Maria Bantz (In preparation). Corpus of Age-related Voice Disguise (AVOID) [speech corpus], Kielipankki - The Language Bank of Finland URL: <http://urn.fi/urn:nbn:fi:bn-201806062>.

License(s)
Rosa González Huartamäki [Ri](#)
University of Eastern Finland [Ri](#)
Tomi Kinnunen [Ri](#)
Distribution rights holders:
University of Jyväskylä [Ri](#)

RPI Holder
Stefan Wiener [Ri](#)
Tomi Kinnunen [Ri](#)
Mäe Sahoulian [Ri](#)
Rosa González Huartamäki [Ri](#)
Maria Bantz [Ri](#)
Ville Huutaniemi [Ri](#)

Contact Person(s)
Tomi Kinnunen [Ri](#)
Rosa González Huartamäki [Ri](#)

Bilingual text corpus

Languages
English (2 Sentences)
Finnish (15 Sentences)

Uniquely
Linguistic type: Bilingual
Multi-linguality type: Other (The Finnish translations of the stories "Harrow Passages" and "North Wind and the Sun", and two English sentences from TIMTE1) (5x1, 5x2)

Size
13 Sentences

Metadata

Created: 05/09/2018
Last Updated: 08/09/2021
Revision: Link to resource group page above

Metadata Creator
Who did create this?
Metka Lesniec [Ri](#)
Documentation
How to cite: <https://www.kielipankki.fi/>
Resource group page: <http://urn.fi/urn:nbn:fi:bn-201806062>
License - Lisenssi: <https://www.kielipankki.fi/>

Document Type: Manual
Object identifier: <http://urn.fi/urn:nbn:fi:bn-201806062>
Publisher: <http://urn.fi/urn:nbn:fi:bn-201806062>
Document Language: Finnish



Tools

www.kielipankki.fi/tools

Start	Name	Description	Instructions	Install	Info	Administrator	Service level
	Korp	A web-based concordance tool that can be used for corpus queries based on morphosyntactic analysis and various other features.	Instructions				A
Download	Download service	Download certain corpora.					A
META-SHARE	META-SHARE	Metadata repository of all the language resources at the Language Bank of Finland.					A
Myilly	Myilly	Versatile data analysis platform with interactive visualizations and workflows.	Instructions				C
Sanat	Sanat	A platform for publishing lexica and word lists.					B
FinTag	Finnish Tagtools	A part-of-speech and morphology tagger and a named entity recogniser for Finnish.		Install Use via Docker			A
Demo	Demo tools at the Language Bank of Finland	Demos of tools that are in development at the Language Bank of Finland: FinTag and FINER, FinSentiment, FinWordNet, HFST POS taggers, HFST morphological analyzers, Lemmamatch, etc. (In Finnish)					C
	WebAnno	Text annotation tool.	User Guide	Standalone installation			A
	Signbank	Lexical database of Finnish Sign Language.					A

<https://docs.csc.fi/apps/>

Language Research and Other Digital Humanities and Social Sciences

- [eBay's tsv-tools](#) Utilities for manipulating large tabular data files
- [finnish-parse](#) Dependency Parser for Finnish
- [Finnish Tagtools](#) Finnish Tagtools
- [HFST](#) Helsinki Finite-State Transducer Technology
- [HFST-fi](#) Helsinki Finite-State Technology for Finnish
- [HFST-sv](#) Helsinki Finite-State Technology for Swedish
- [Kaldi](#) Kaldi Speech Recognition Toolkit
- [UDPipe](#) UDPipe Kielipankki version



Data on HPC

```
[matthies@puhti-login15 kielipankki]$ cd /appl/data/kielipankki
[matthies@puhti-login15 kielipankki]$ ls
acquis          Kalevala        sfnet
admin           kksxml          SNC1
amph            kra             suomi24-2001-2017-vrt-v1-2
bert_models    lehdet-vrt-v2  suomi24-2018-2020-vrt-beta
dspcon2013-2016-dl mrc-uhlcs      susanne
Eduskunta      nlfcl-fi-vrt   wikipedia-fi-2017-src
FNC1           nlfcl-sv-vrt   wordnet
hcs-a-v2-dl    ota             words
hcs-na-v2      oulu           ylilauta
[matthies@puhti-login15 kielipankki]$ █
```



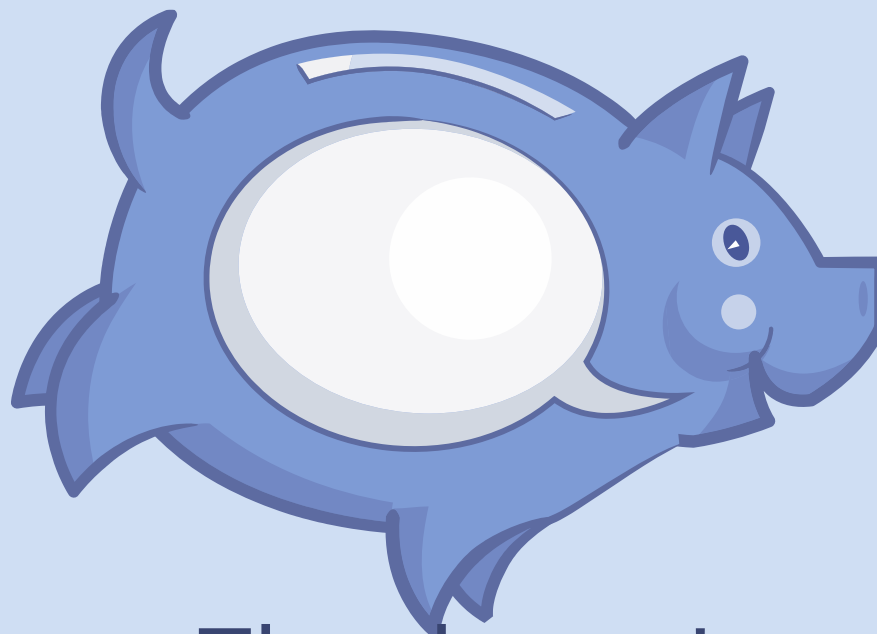
CLARIN ERIC

- 22 Members
- 2 Observers
- 70 Centres
- Metadata Search: <https://vlo.clarin.eu>
- Metadata Standards: CMDI
- Federated Content Search
- Persistent Identifiers: Handles, URNs
- Federated Login
- License Framework (PUB, ACA, RES)



KIELIPANKKI

The Language Bank of Finland



Thank you!
www.kielipankki.fi