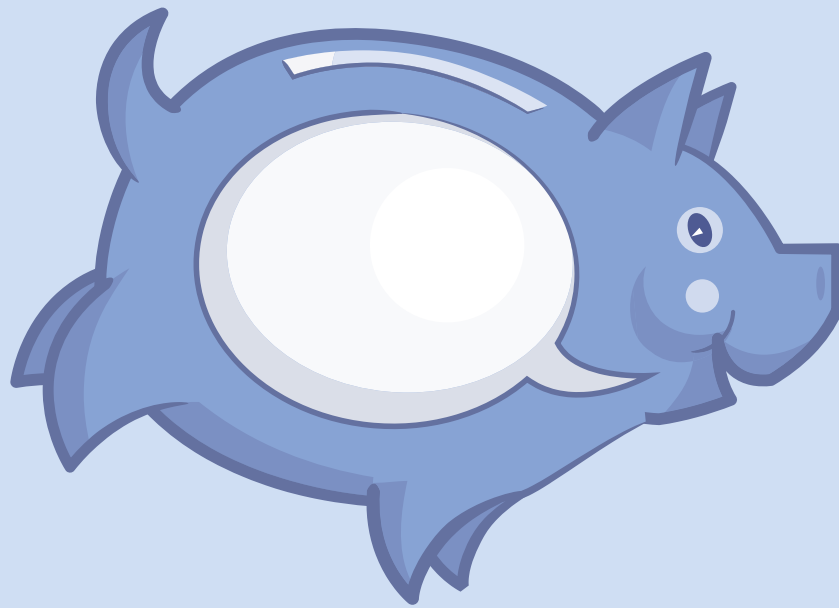




KIELIPANKKI
The Language Bank of Finland

Kielipankki and Korp

Mietta Lennes and Jyrki Niemi, FIN-CLARIN / University of Helsinki
Korp & Språkbanken workshop 17.10.2018

This document is licensed under Creative Commons Attribution 4.0.



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:

Hae



Researcher of the Month: Annika Liiti

News

- Researcher of the Month: Annika Liiti (13.9.2018)
- Researcher of the Month: Maria Huttu-Hiltunen (6.8.2018)
- Researcher of the Month: Mielita Lennes (9.7.2018)
- Researcher of the Month: Maximilian Murmann (11.6.2018)
- Researcher of the Month: Ivana Kováčová (7.5.2018)



FIN-CLARIN Partners

- University of Helsinki
- CSC – IT Center for Science

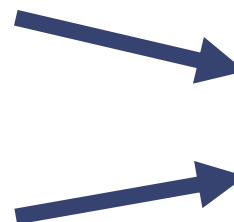


Coordination and access to large **centrally acquired resources and tools**

- KOTUS – Institute for the Languages of Finland
- Aalto University
- University of Eastern Finland
- University of Jyväskylä
- University of Oulu
- University of Tampere
- University of Turku
- University of Vaasa



Access to **resources and tools developed locally** by individual researchers or research groups



CLARIN

Common Language Resources and Technology Infrastructure



Resources	2018	2022
<i>Text</i>		
Magazines and newspapers 1770- (NLF and Web publ.)	12 Gw	
Social media and similar sources 2000- (Suomi24, Ylilauta, ...)	4 Gw	
Literature and manuscripts (Gutenberg, Fennica, archives)	60 Mw	70 Mw
<i>Speech</i>		
News broadcasts (YLE)		10000 h
Video sessions from the Finnish Parliament 2008-2016	500 h	1000 h
Dialect and everyday speech (Kotus, Turku)	500 h	1000 h
Sign language resources (Aalto, Kuurojen liitto)	20 h	500 h
<i>Multilingual and Other Resources</i>		
Multilingual Resources (EuroParl, laws, Bible, subtitles, ...)	3 Gw	10 Gw
Learner's resources (Oulu, Jyväskylä, Kotus, Aalto)	2 Mw	5 Mw
Open source lexicons and terminologies (Helsinki, Tromssa)	300 Kw	400 Kw

total:
~19 billion
words

Corpora deposited in the Language Bank

Location

The corpora in the Language Bank may be accessible via a web interface ([Korp](#) and [LAT](#)) or by using command line tools on the Language Bank's software server. Some corpora can also be downloaded via the [Download service](#).

License

Some corpora are directly available , some may be accessed by signing in  or by applying for individual access rights . Protected corpora can be accessed following the link  in the *Apply* column.

- [How to get access to the language resources of the Language Bank of Finland](#)

Cite

Corpora-specific reference instructions are available by clicking the quote  link.

Etsi:

Abbreviation	Name and metadata	License	Apply	Location	Help	Cite
acquis-ftb3	The Finnish Sub-corpus of the JRC-Acquis Multilingual Parallel Corpus			Korp		
Agricola	The Morpho-Syntactic Database of Mikael Agricola's Works			Korp		
aku-egg	Speech and EGG (Electroglottography) Simultaneous Recordings			LAT		
amph	amph-Corpus			Download		
ArkiSyn-korp	ArkiSyn Database of Finnish Conversational Discourse, Helsinki Korp Version			Korp		
BeserCorp	The Corpus of Beserman Udmurt			Korp		
ceal-o	Classics of English and American Literature in Finnish, Sentences and Paragraphs in the Original Order			Korp		



The Suomi 24 Sentences Corpus (2016H2)

505  18

View resource name in all available languages

Suomi24-korp-2016H2

Persistent Identifier of this resource: <http://urn.fi/urn:nbn:fi:lb-2017021505>

Access location: <http://urn.fi/urn:nbn:fi:lb-2015120401>

The corpus is available in Kielipankki (the Language Bank of Finland (korp.csc.fi), <http://urn.fi/urn:nbn:fi:lb-2015120401> (sub-corpora 1/10-10/10).

The corpus contains all the discussion forums of the Suomi24 online social networking website from 1st January 2001 to 24th September 2016 available... [Read More](#)

View resource description in all available languages

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

Persistent ID (for references & citation)

Link to the current access location

Distribution

Availability

Available - Unrestricted Use

Licence

CC - BY - NC

Restrictions: Attribution

Attribution Details: See Documentation section.

Distribution Access/Medium: Accessible Through Interface

Licensors:

Aller Media Oy 

Distribution rights holders:

University of Helsinki 

IPR Holder

Aller Media Oy 

Contact Person

User support FIN-CLARIN 

text

Monolingual text corpus

Languages

Finnish

Linguality

Linguality type: Monolingual

Size

2,663,114,497 Tokens

Modalities

Written Language

Time Coverage

1.1.2001-24.9.2016

Metadata

Created: 09/17/2015

Last Updated: 09/17/2015

Metadata Language: English, Finnish (en, fi)

Metadata Creator

Imre Bartis 

Version

Version: 2016H2

Relation

Related Resource: The Suomi 24 Corpus

<http://urn.fi/urn:nbn:fi:lb-2015120401>

Relation Type: HasMetadata

Related Resource: The Suomi 24 Corpus (2016H2) <http://urn.fi/urn:nbn:fi:lb-2015120401>

Relation Type: IsVariantFormOf

Related Resource: The Suomi 24 Sentences Corpus (2015H2) <http://urn.fi/urn:nbn:fi:lb-2015120401>

Relation Type: IsNewVersionOf

Documentation

Attribution Details

<https://www.kielipan...>



The Suomi 24 Sentences Corpus (2016H2)

505  18

View resource name in all available languages

Suomi24-korp-2016H2

Persistent Identifier of this resource: <http://urn.fi/urn:nbn:fi:lb-2017021505>

Access location: <http://urn.fi/urn:nbn:fi:lb-2015120401>

The corpus is available in Kielipankki - the Language Bank of Finland (korp.csc.fi), <http://urn.fi/urn:nbn:fi:lb-2015120401> (sub-corpora 1/10-10/10).

The corpus contains all the discussion forums of the Suomi24 online social networking website from 1st January 2001 to 24th September 2016 available... [Read More](#)

View resource description in all available languages

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

Citation instructions

[\[suomeksi\]](#) [\[in English\]](#)

Reference instructions: Suomi24-korp-2016H2

Please cite the language resource as follows:

Aller Media Ltd. (2014). *The Suomi 24 Sentences Corpus (2016H2)* [text corpus]. Kielipankki. Retrieved from <http://urn.fi/urn:nbn:fi:lb-2017021505>

[\[Bibtex\]](#) [\[Zotero\]](#)

through interface

Licensors:

[Aller Media Oy](#) 

Distribution rights holders:

[University of Helsinki](#) 

IPR Holder

[Aller Media Oy](#) 

Contact Person

[User support FIN-CLARIN](#) 

Written Language

Time Coverage

1.1.2001-24.9.2016

Related Resource: The Suomi 24 Corpus

<http://urn.fi/urn:nbn:fi:lb-2015120401>

Relation Type: HasMetadata

Related Resource: The Suomi 24 Corpus

(2016H2) <http://urn.fi/urn:nbn:fi:lb-2015120401>

Relation Type: IsVariantFormOf

Related Resource: The Suomi 24 Sentences

Corpus (2015H2) <http://urn.fi/urn:nbn:fi:lb-2015120401>

Relation Type: IsNewVersionOf

Documentation

Attribution Details

<https://www.kielipan...>



Speech corpora in LAT (lat.csc.fi): Links to the Annex viewer from Korp

 **Annex 1.6.42702** manual ? embed

☒ Show tooltips

Compact

Spacious

user: anonymous login

Text

Grid

Subtitle

Waveform

Timeline

Combined

Video display

min



Information

min

General Session Technical

Resource: reitti_a-siipeen.eaf
Media file: reitti_a-siipeen.mp4
Elapsed time: 00:00:07:827

Selected chunk:
Begin time: 00:00:07:827
End time: 00:00:12:205

Text:
voit sä neuvoo mulle jos mä seison tual
fabianinkadulla .fp ja mä haluun tulla sisään
mahollisimman nopeesti mul

Mini Data Frame

tänne sun huaneesee o hirvittävä vaikee löytää
voit sä neuvoo mulle jos mä seison tual
fabianinkadulla .fp ja mä haluun tulla sisään
mahollisimman nopeesti mul on kiire mitä mä
eli okei mä oon tolla pualella selvä sitte mitä s
o ne portaat jois o ne ikkunat oikeelle .bri aa
eli beesipi on tuolla ja nyt me ollaa aasiivessä
mm hyvä ja sit mä käänny vasemmalle ja .fp a
nythän se o ihan täysin selvää kiitos mm ehkä

Tier: TA-utterance

Font size: 14

Play selection

Clear selection

Create bookmark

|< >|

<< >>

< >

+ -

☒ Play screen by screen

☐ Play continually

Timeline





Download service: www.kielipankki.fi/download

KIELIPANKKI
The Language Bank of Finland

LATAUKSET
DOWNLOADS

HOME UP

Location: /download/ Logged in as: (none)

Name	Size	Description
AMPH/	-	amph-Corpus
BYU/	-	The BYU corpora at Kielipankki - The Language Bank of Finland
CEAL/	-	CEAL corpus
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
finka/	-	Raja-Karjalan korpus
finnish-tagtools/	-	Finnish Tagtools
FNC1/	-	Kansalliskirjaston lehtikokoelman suomenkieliset n-grammit 1820-2000
FTC/	-	Finnish Text Collection
HCS/	-	Helsinki Corpus of Swahili 2.0
helpuhe/	-	The Longitudinal Corpus of Finnish Spoken in Helsinki
hfst-morphologies/	-	HFST morphologies for various languages
italian-letters/	-	Italian Letters from the 16. Century
KKS/	-	Karjalan kielen sanakirja (XML)
LAS2/	-	Advanced Finnish Learners' Corpus



Mylly – “The Mill”

interactive web services by FIN-CLARIN

Import query results over Korp API

Analysis tools

- Kielipankki
 - Korp search
 - Korp data**
 - Relation algebra
 - Simple statistics
 - Advanced statistics
 - Export formats
 - Syntactic analysis
 - Morphological analysis
 - Speech recognition
 - Preprocessing
 - Finite-State Technology
 - Finite-State Transducers
 - Job management

Visualisation

Showing 3609 rows of 3609 and all 12 columns

kMmatch	kMsen	kMtok	lemma	deprel	ref	aid	pos	nertag	word	msd	dephead
0	0	0	silloin	advmod	1	a12594	Adv	-	Silloin	CASECHA...	3
0	0	1	täytyä	aux	2	a12595	V	-	täytyy	PRS_Sg3 V...	3
0	0	2	uskaltaa	ROOT	3	a12596	V	-	uskaltaa	NUM_Sg C...	0
0	0	3	nostaa	xcomp	4	a12597	V	-	nostaa	NUM_Sg C...	3
1	0	4	kissa	doobj	5	a12598	N	-	kissa	NUM_Sg C...	4
0	0	5	pöytä	nommod	6	a12599	N	-	pöydälle	NUM_Sg C...	4
0	0	6	arvoisa	amod	7	a12600	A	-	arvoisa	NUM_Sg C...	8
0	0	7	puhemies	doobj	8	a12601	N	-	puhemies	NUM_Sg C...	4
0	0	8	.	punct	9	a12602	Punct	-	.	-	3
1	1	0	ja	advmod	1	a17480	Adv	-	Ja	CASECHA...	5
0	1	1	sitten	advmod	2	a17481	Adv	-	sitten	-	3
0	1	2	vihonviimei...	acomp	3	a17482	A	-	vihonviimei...	NUM_Sg C...	5
0	1	3	lopuksi	advmod	4	a17483	Adv	-	lopuksi	-	5
0	1	4	haluta	ROOT	5	a17484	V	-	haluan	PRS_Sg1 V...	0
0	1	5	nostaa	xcomp	6	a17485	V	-	nostaa	NUM_Sg C...	5
0	1	6	se	doobj	7	a17486	Pron	-	sen	SUBCAT_D...	6
1	1	7	kissa	poss	8	a17487	N	-	kissan	NUM_Sg C...	9
0	1	8	pöytä	nommod	9	a17488	N	-	pöydälle	NUM_Sg C...	6
0	1	9	joka	rel	10	a17489	Pron	-	jota	SUBCAT_R...	12
0	1	10	tämä	nommod	11	a17490	Pron	-	tässä	SUBCAT_D...	12
0	1	11	kiertää	rcmod	12	a17491	V	-	kiertetään	PRS_Pe4 V...	9
0	1	12	kuin	comparator	13	a17492	C	-	kuin	SUBCAT_CS	14



Language Bank support



LANGUAGE BANK ACCESS CORPORA TOOLS ORGANIZATION SUPPORT

SUOMEKSI PÅ SVENSKA

Korp user guide

Korp is a Web-based tool that allows its user to search for keywords in text corpora (typically grammatically parsed) and to generate concordances. Korp gives its users access to extensive collections of texts in Finnish and Finland Swedish. Many of these corpora can be accessed without logging in whereas others are only accessible to logged-in users. In some cases individual access rights are also required ([instructions for applying](#)).

Please note that this manual has been written specifically for the version of Korp used by the [Language Bank of Finland](#) and may not be fully applicable to other versions. Korp is developed at [Språkbanken](#) of the University of Gothenburg in Sweden, and their Korp site contains mostly Swedish-language materials. The [Norwegian version of Korp](#) contains corpora in Norwegian and Saami. [Instructions for Korp as used by the Swedish Språkbanken](#) (in Swedish only – note that these instructions may not be entirely applicable to the Finnish version of Korp).

Start Korp: <https://korp.csc.fi/#?lang=en>

General information

Choosing the environment

Selecting a corpus

Choosing the language



Kielipankki on YouTube

Introduction to Korp, part 1:

<https://youtu.be/YkIHHRbbA40>

(part 2 is forthcoming...)



KIELIPANKKI
The Language Bank of Finland

FIN-CLARIN

Online courses

Corpus Linguistics and Statistical Methods (5 cr)

Word attributes

- baseform: kieli
- baseform (compound boundaries): kieli
- part-of-speech: noun
- msd: NUM_sg/CASE_Nom
- dependency relation: nominal subject
- word position in a name: outside (O)
- original thread
- original message
- Show Dependency Tree
- Näytä dependenssipuu



Introduction to Speech Analysis (5 cr)

tänne sun huon
no niin varmaa
sit sa neuvoo
tököön mähöi
tää mä
täninkatu ei
sa sa tuut ni
okey sa ooi
lvä



Corpus Clinic (5 cr)

A	B	C	D	E	F	G	H	I
1	match	sen	tok	deprehd	med			
2	0	0	0	2 SUBCAT_Pu	[NUMP_sg/CASE_Nom]	terma	lex	ref
3	0	0	1	3 NUMP_P	[CASE_Nom]	polka	[lexica_pu 1]	5.54
4	0	0	2	0 PROS_Agt	[VALOR_Agt]	valon	[lexica_pu 1]	2.44
5	0	0	3	3 NUMP_sg	[CASE_Nom]	juoma	[lexica_pu 1]	4.44
6	0	0	4	4		reuna	[lexica_pu 1]	4.44
7	0	0	5	4 SUBCAT_CC				0.44
8	0	0	6	0 SUBCAT_Dum	[NUMP_sg]			0.44
9	0	0	7	0 PROS_Agt	[VALOR_Agt]			0.44





Students frequently ask...

- How do I know whether corpus X is annotated or not?
- How do I find a linguistic construct of the type XYZ?
 - (General answer: look for at least one example, check out the word properties and use them in Extended search)
- Why do I "always" see a dead Korp?



Researchers frequently ask...

- (Do you have corpora of type X?)
- Which annotation system was used in corpus X?
- How do I obtain a list of the (10/20) most frequent words(/nouns/verbs/other) in corpus X / language Y?
- How to obtain a random sample?
- I want to correct/ I have corrected some annotations to a corpus. How can they be **inKorporated**?



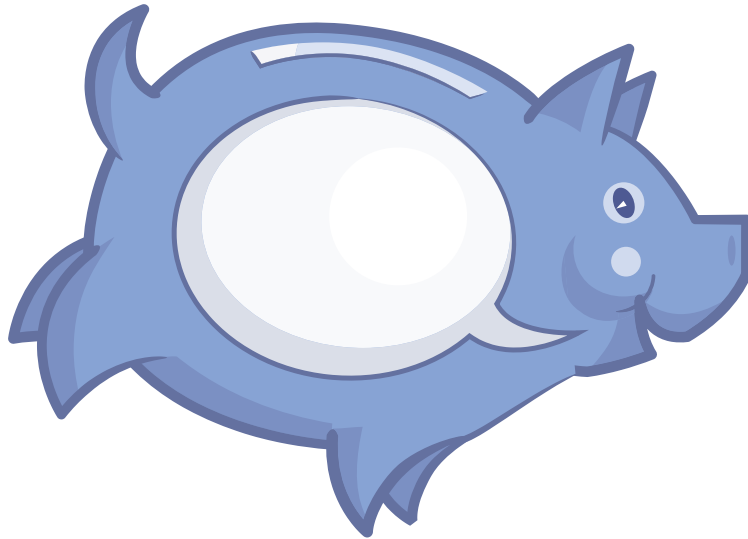
Current Korp.csc.fi

- Based on Korp frontend 5.0.10, backend 2.8 (almost two years old)
- A large number of local, relatively minor modifications
 - Shibboleth authentication
 - KWIC results download (export)
 - Standardized corpus information
 - Should be ported when upgrading Korp, complicating upgrades



Future work

- Unify Korp corpus attributes
 - Add Universal Dependencies 2.0 parses alongside the current annotations
- New and updated corpora
 - Enhance the corpus conversion pipeline, previously rather ad hoc
- Update Korp to the most recent version and port our modifications to it
 - Some modifications are not needed any more
 - Integration to the common Korp codebase?
- Add geographic metadata to enable the new map
- Try to add some new features to Korp
 - Enhance the parallel mode
 - Enhance KWIC download formats
 - Add some options to presentation of KWIC results?



Thank you!

www.kielipankki.fi

General support
fin-clarin@helsinki.fi

Technical support
kielipankki@csc.fi