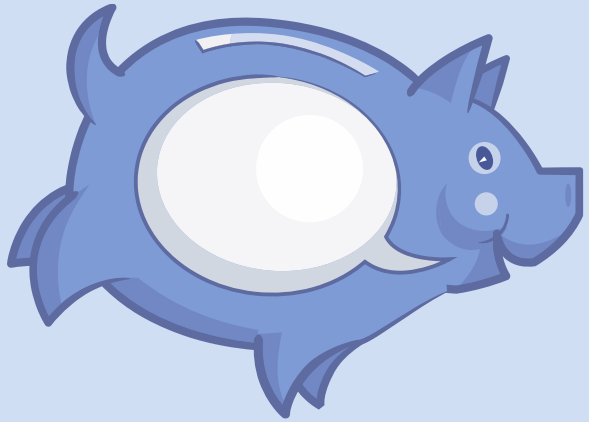


KIELIPANKKI
The Language Bank of Finland

Finding, using and sharing research data via

Kielipankki – The Language Bank of Finland

Mietta Lennes
mietta.lennes@helsinki.fi

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:

Haku ... Hae

Researcher of the Month: Heini Kallio

News

- Researcher of the Month: Heini Kallio (12.3.2026)

Users of the Language Bank

- Researchers from all fields welcome!
- Many corpora are available even without signing in.
- FIN-CLARIN can help you in storing and distributing your own resource.

https://www.kielipankki.fi/language-bank/researcher-of-the-month-archive/

The Language Bank of Finland Researchers of the Month

Suomeksi

Do you know researchers who use the Language Bank of Finland and who might be good candidates for Researcher of the Month? Would you be one of them? Inform us!

Etsi:

Published	Researchers	Corpora/Tools	Publications
2026.02	Kalle Lahtinen	FinnAffect, puhelahjat, helpuhe, tampuhe	2025a, 2025b
2026.01	Atte Huhtala	KLK-fi, suomi24, la-murre, sapu, DMA, SKN, agricola	2023, 2022
2025.12	Satu Siltaloppi	cfsts	2025, 2023a, 2023b, 2019
2025.11	Krista Ojutkangas	suomi24, ftc, klk-fi, klk-sv, vks, vnsk, ark-isyn	2025, 2023a, 2023b, 2020, 2017, 2013
2025.10	Dejan Porjazovski	fi-parliament-asr, puhelahjat	2024, 2023a, 2023b, 2021, 2020
2025.09	Inka Rantakallio	femrap	2025a, 2025b, 2021
2025.08	Idastiina Valtasalmi	lehdet90ff, ylenews	2024, 2023
2025.07	Rea Peltola	la-murre, dma, vnsk, ylenews, suomi24, lehdet90ff, skk, arkisyn	2023, 2021
2025.06	Jörg Tiedemann	Opus-Korp	2024, 2023a, 2023b, 2022

Search the Language Bank Portal:

Haku ... Hae

Researcher of the Month: Heini Kallio

News

- Researcher of the Month: Heini Kallio (12.3.2026)
- New resources: Yle Finnish News Archive 2022-2024 in VRT format (10.3.2026)
- A new SKS publication: Sanovat syntaksiksi on Data-based approaches to Finnish dialects (27.2.2026)
- Researcher of the Month: Kalle Lahtinen (13.2.2026)
- Researcher of the Month: Atte Huhtala (21.1.2026)

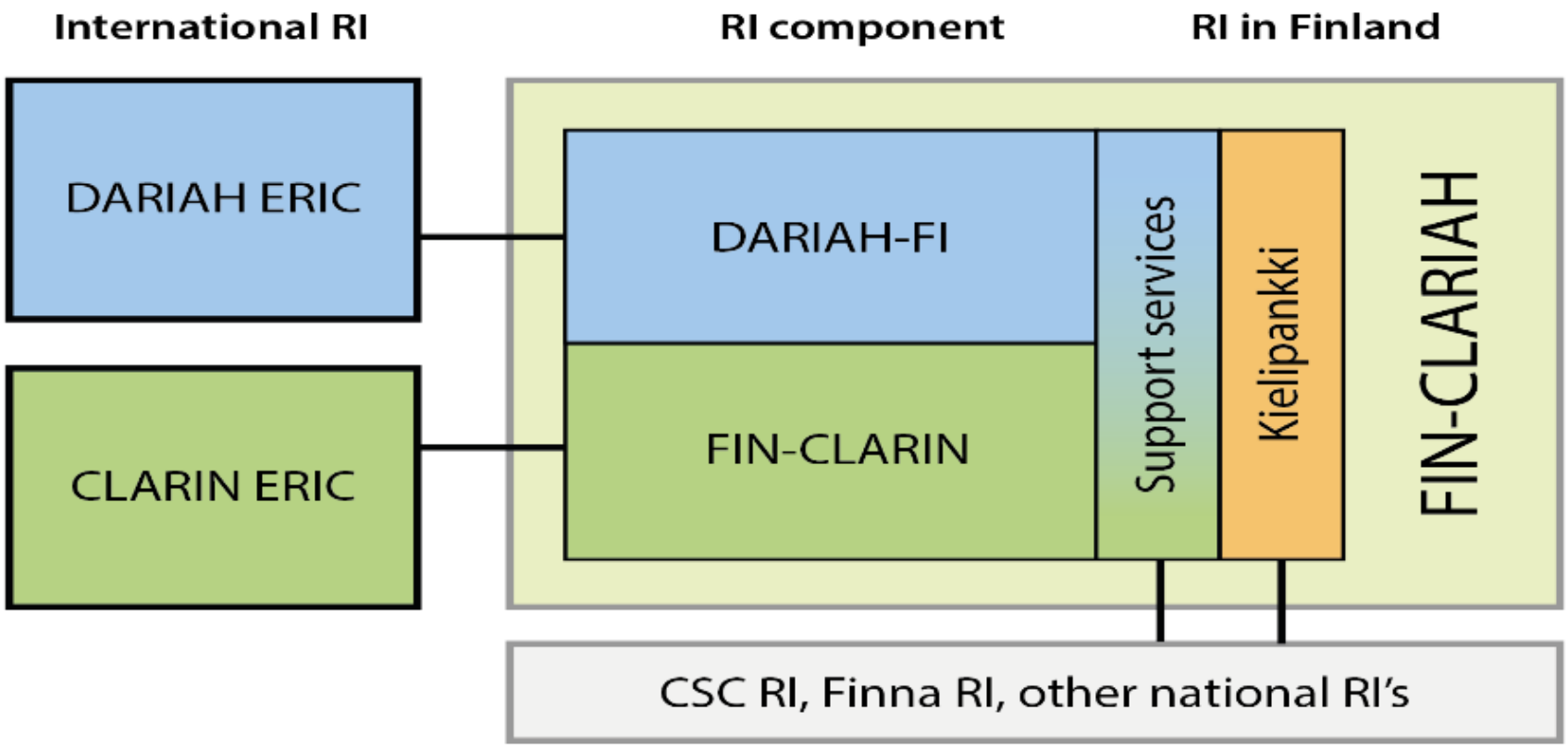
Findable

Accessible

FAIR data

Interoperable

Re-usable


International RI
RI component
RI in Finland
DARIAH ERIC
CLARIN ERIC
DARIAH-FI
FIN-CLARIN
Support services
Kielipankki
FIN-CLARIAH
CSC RI, Finna RI, other national RI's

CLARIN ERIC

European Research Infrastructure Consortium
founded on February 29, 2012

https://www.clarin.eu

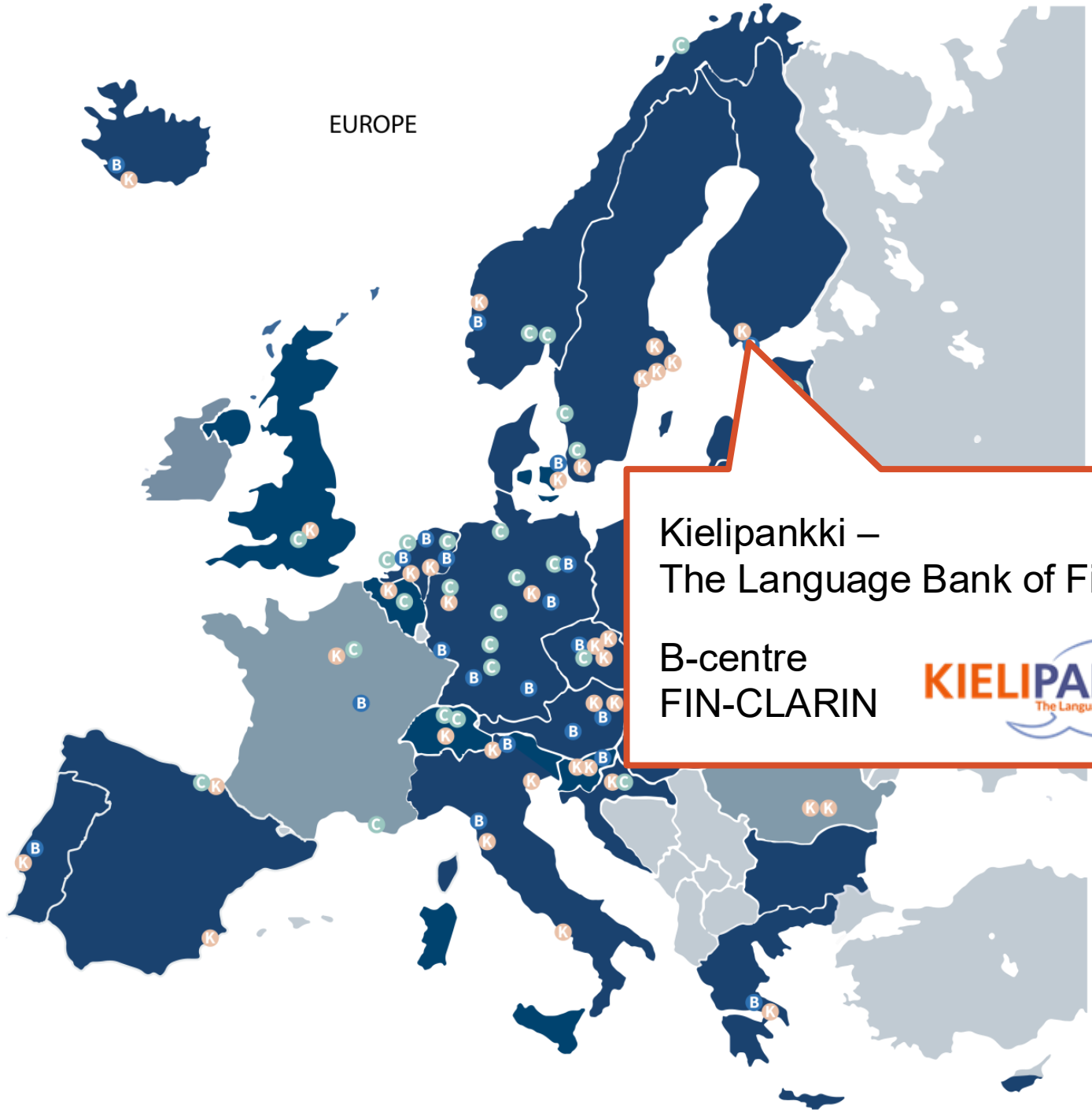
Other
B centres,
C centres and
K centres at
27* national consortia

International cooperation
and sharing of resources
for Humanities and Social Sciences

**17.3.2026*

CLARIN

- ERIC members
- Observers
- Countries with participating centres
- (B) Centre Providing Data
- (C) Centre Providing Metadata
- (K) Knowledge Centre



Kielipankki –
The Language Bank of Finland

B-centre
FIN-CLARIN

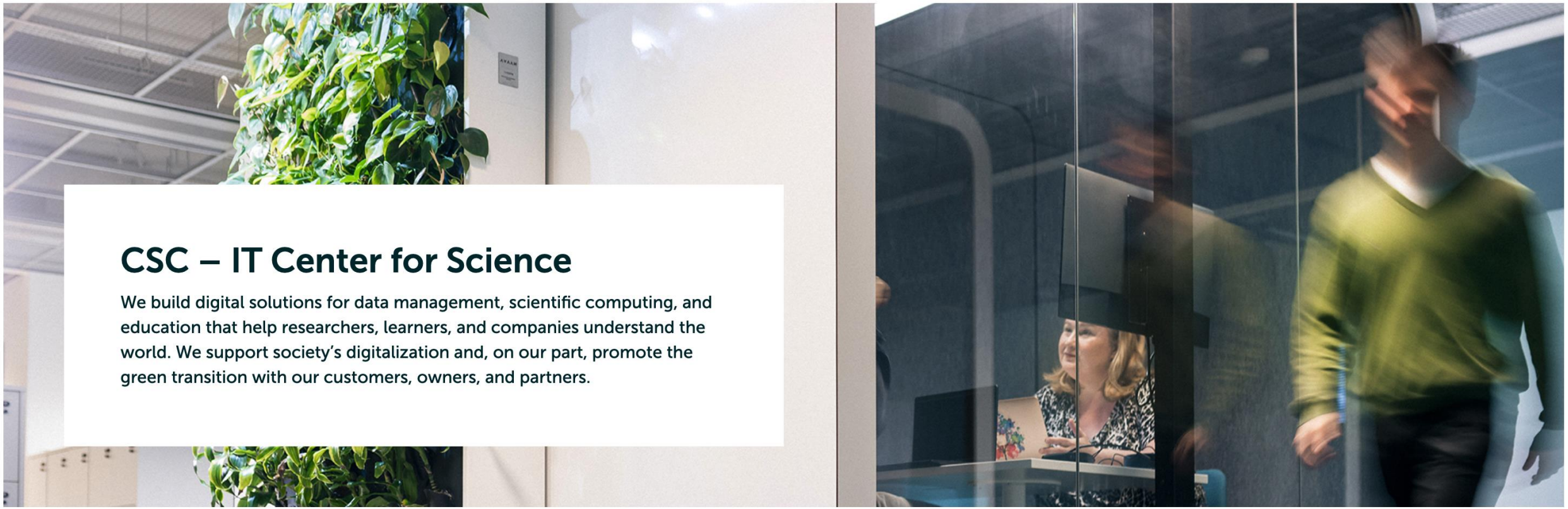
KIELIPANKKI
The Language Bank of Finland

https://csc.fi/

CSC

In English | Search | Other sites | Service Catalog

Front page | Our expertise | About us | Careers | Training | News



CSC – IT Center for Science

We build digital solutions for data management, scientific computing, and education that help researchers, learners, and companies understand the world. We support society's digitalization and, on our part, promote the green transition with our customers, owners, and partners.

CLARIN Virtual Language Observatory

Welcome to the VLO!

Use the **search bar** below to start searching through hundreds of thousands of language resources, or continue to browse everything and use **facets** to narrow down to your area of interest or discover new resources.

See all records | Take a quick tour

Search through 911,067 records

Showing all records (574,361 results)

Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

- Language
- Collection
- Resource type
- Modality
- Format
- Keyword

<< < 1 2 3 4 5 6 7 8 9 10 > >>

Nganasan Spoken Language Corpus (NSLC)

176 1

(Part of Hamburger Zentrum für Sprachkorpora (HZSK))

The Nganasan Spoken Language Corpus (NSLC) has been created as part of Corpus based grammatical studies on Nganasan project (supported by the German Research Grant; WA3153/2-1). The Spoken Nganasan Corpus contains the same text samples in at least three languages: The original text in Nganasan with translations mostly …

Nganasan | Russian

Landing page for this record

EXMARaLDA Demo corpus

(Part of Hamburger Zentrum für Sprachkorpora (HZSK))

A selection of short audio and video recordings in various language

https://vlo.clarin.eu

Resource families

Introduction

The aim of the CLARIN Resource Families initiative is to provide a user-friendly overview per data type of the available language resources in the CLARIN infrastructure aimed at the needs of researchers from digital humanities, social sciences and human language technologies. The overviews are meant to facilitate comparative research and the listings are sorted by language.

The listings for each family include the most important metadata and brief descriptions, such as resource size, text sources, time periods, annotations and licences as well as links to download pages and concordancers, whenever available. In addition to the resources found in the CLARIN infrastructure an overview is provided of other existing valuable language resources which have not yet been integrated in the infrastructure.

The listings also provide hyperlinks to other relevant materials such as the thematic CLARIN workshops and tutorials and their accompanying videolectures, as well as a list of key publications on the resources surveyed.

Currently overviews are available of 12 corpora families, 5 families of lexical resources, and 4 tool families. See below. For the possibility to apply for funding for small projects that can help extending the scope of the initiative, see https://www.clarin.eu/content/clarin-resource-families-project-funding.

Corpora

- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora

Lexical Resources

- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools

- Normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Tools for sentiment analysis

Resource families

CMC corpora

Historical corpora

L2 corpora

Manually annotated corpora

Multimodal corpora

Newspaper corpora

Parallel corpora

Parliamentary corpora

Reference corpora

Spoken corpora

Lexica

Dictionaries

Conceptual resources

Glossaries

Wordlists

Tools for normalization

Tools for named entity recognition

Part-of-speech taggers and lemmatizers

Tools for sentiment analysis

https://www.clarin.eu/resource-families

Books and manuscripts

Parallel corpora and other multilingual resources

Newspapers and journals

Lexicons and terminologies

Sessions of the Parliament (videos + transcripts)

Online discussions and other data from the Internet (e.g., Suomi 24, Ylilauta)

Learner language corpora

Sign languages

Lahjoita puhetta – Donate Speech (puhelahjat)

Dialect corpora

See the complete list of corpora at https://www.kielipankki.fi/corpora/

https://www.kielipankki.fi/corpora

Currently available corpora

Search:

Shortname	Name and metadata	License	Location	Cite	Resource group and help	Apply	Publication year	Support level
achemenet-2020-12-korp	Achemenet Babylonian texts - Kielipankki version 2020-12, Korp	PUB	Korp		achemenet		2025	B
agricola-v1-1-korp	The Morpho-Syntactic Database of Mikael Agricola's Works version 1.1, Korp	PUB	Korp		agricola		2020	B
ai2d-rst-v1-1	AI2D-RST: A multimodal corpus of 1000 primary school science diagrams version 1.1	PUB	Download		ai2d-rst		2020	B
aku-egg-dl	Speech and EGG (Electroglottography) Simultaneous Recordings, downloadable version	ACA	Download		aku-egg		2014	B
amph	amph-Corpus	ACA	Download Puhti		amph		2008	B
ArkiSyn-korp	ArkiSyn Database of Finnish Conversational Discourse, Helsinki Korp Version	PUB	Korp		arkisyn		2017	B
AVOID	Corpus of Age-related Voice Disguise (AVOID)	RES	Download		avoid		2019	B
balt-2025-02-korp	BALT: Babylonian Administrative and Legal Texts - Kielipankki version 2025-02, Korp	PUB	Korp		balt		2025	B
BeserCorp	The Corpus of Beserman Udmurt	PUB	Korp		besercorp		2016	B

Citation instructions

amph	amph-Corpus	ACA	Download Puhti	❞	amph		2008
ArkiSyn-korp	ArkiSyn Database of Finnish Conversational Discourse, Helsinki Korp Version	PUB	Korp	❞	arkisyn		2017
AVOID	Corpus of Age-related Voice Disguise (AVOID)	RES	Download	❞	avoid		2019
balt-2025-02-korp	BALT: Babylonian Administrative and Legal Texts - Kielipankki version 2025-02. Korp	PUB	Korp	❞	balt		2025
BeserCorp							2016
ccmh-src							2021

[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. (2019). *Corpus of Age-related Voice Disguise (AVOID)* [data set]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2018060621

Show: [Bibtex] [Zotero]

Search for references to the language resource in Google Scholar

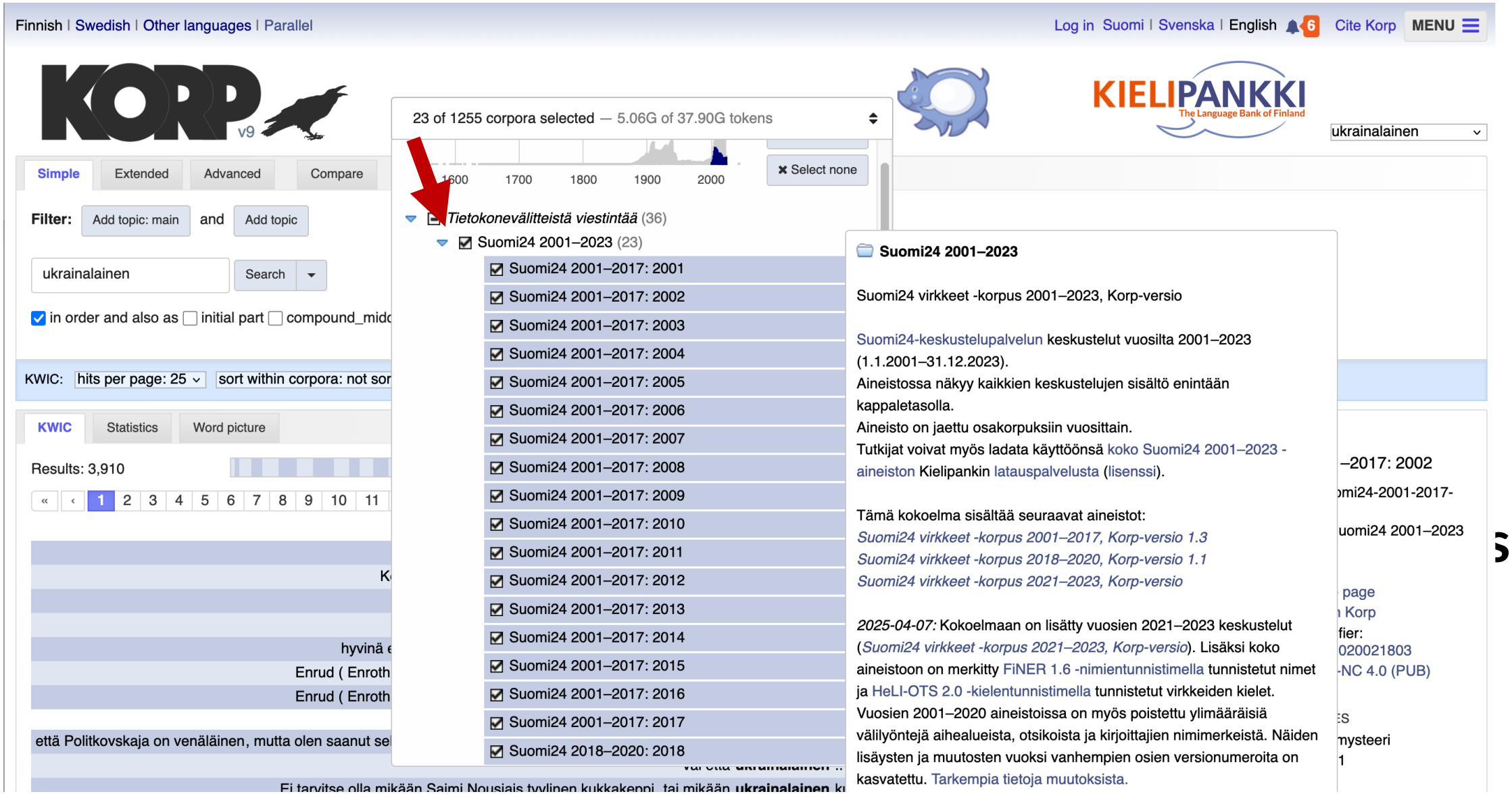
KIELIPANKKI
The Language Bank of Finland

Accessing resources

FIN-CLARIN

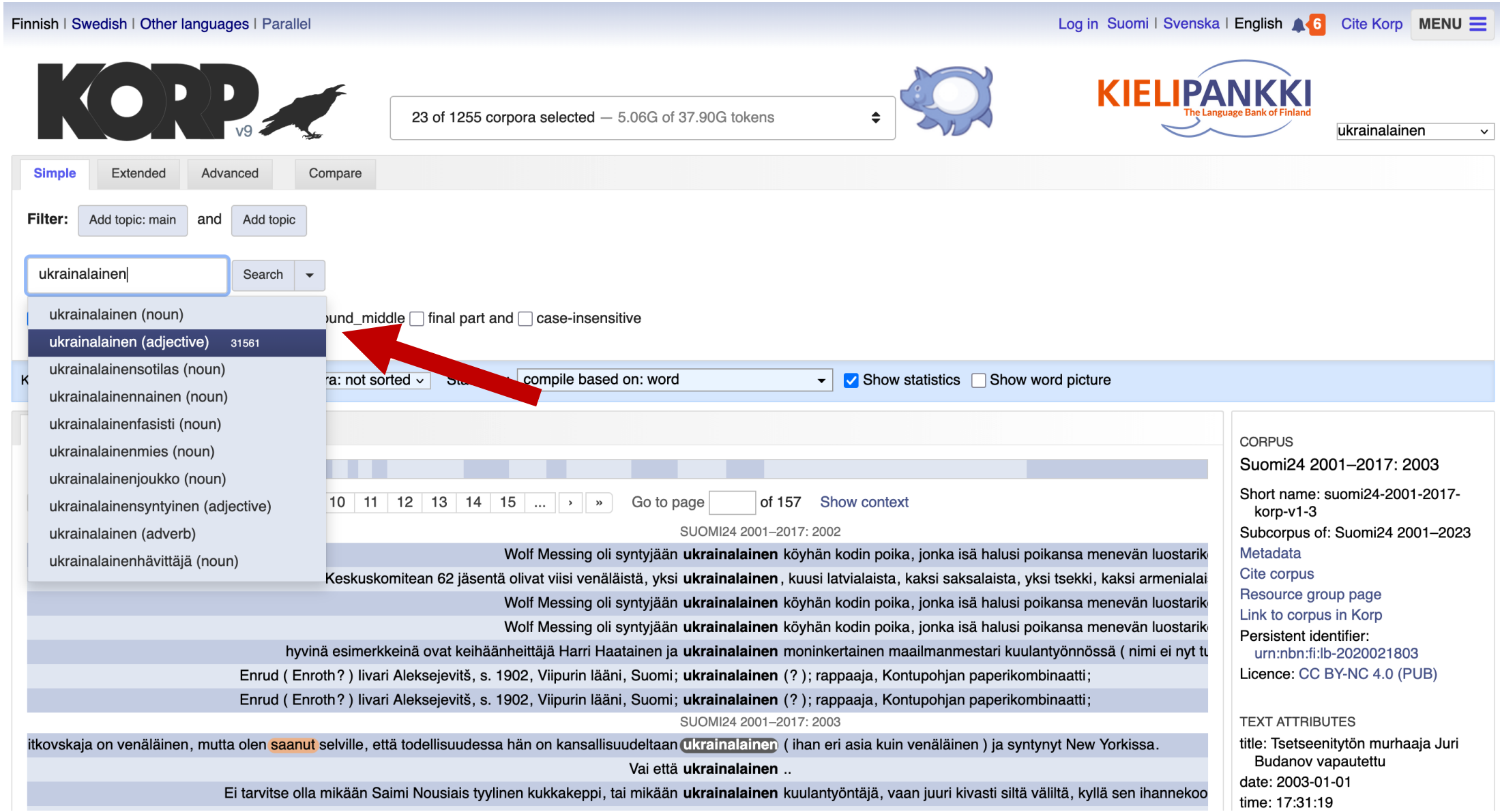
The Suomi24 Sentences Corpus 2021-2023, Korp version

https://www.kielipankki.fi/korp/



The Suomi24 Sentences Corpus 2021-2023, Korp version

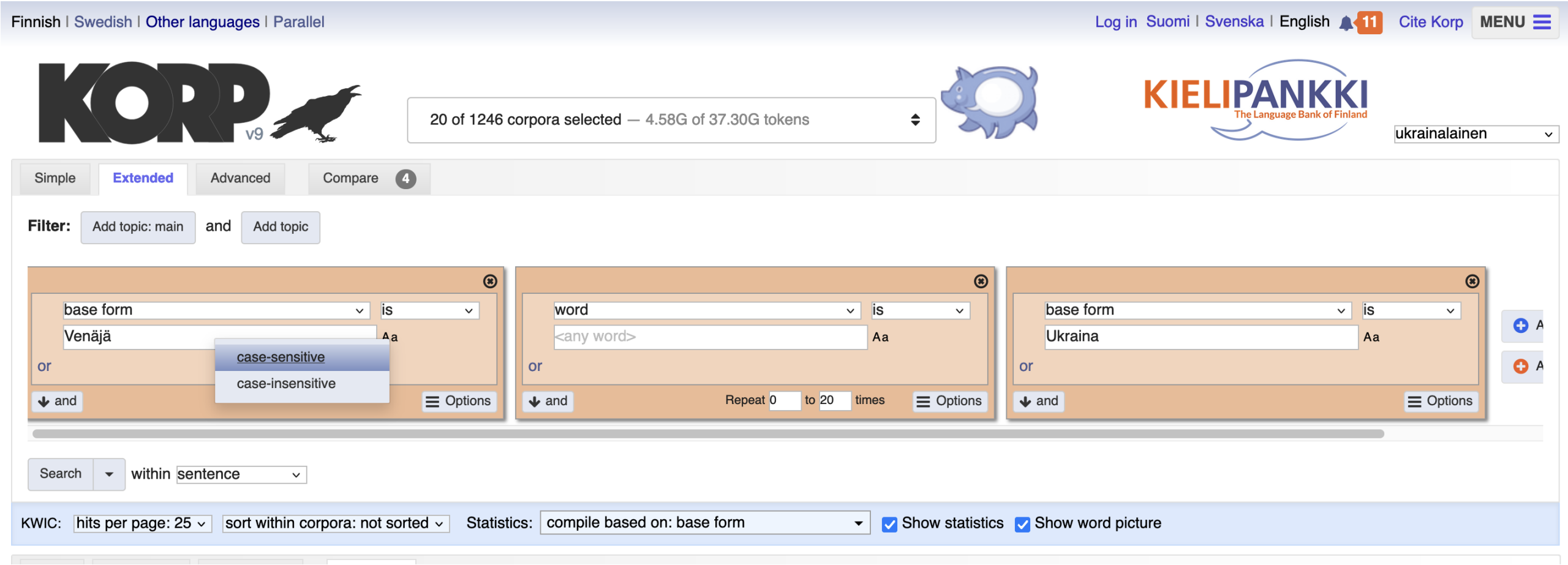
https://www.kielipankki.fi/korp/



The Suomi 24 Sentences Corpus 2001-2020, Korp version

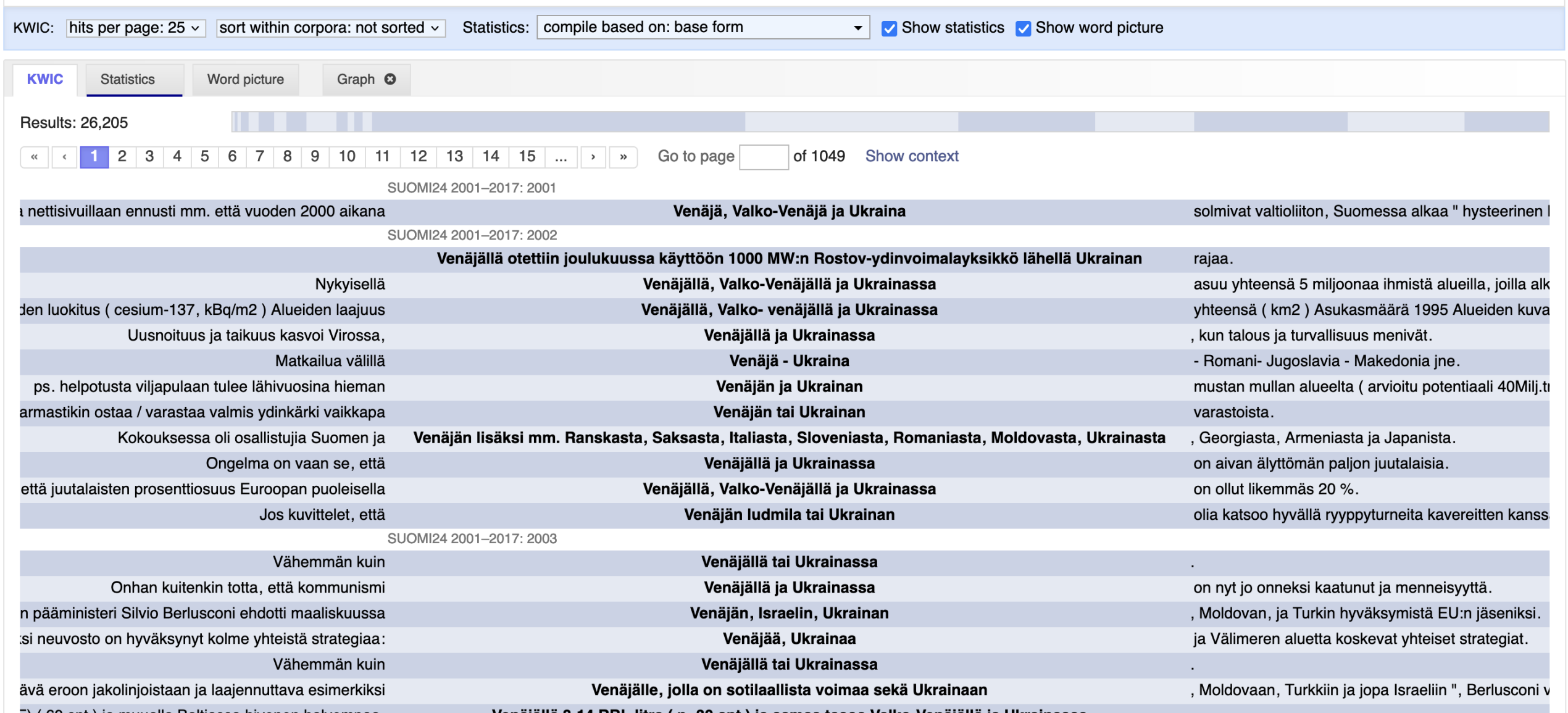
https://www.kielipankki.fi/korp/

Extended search

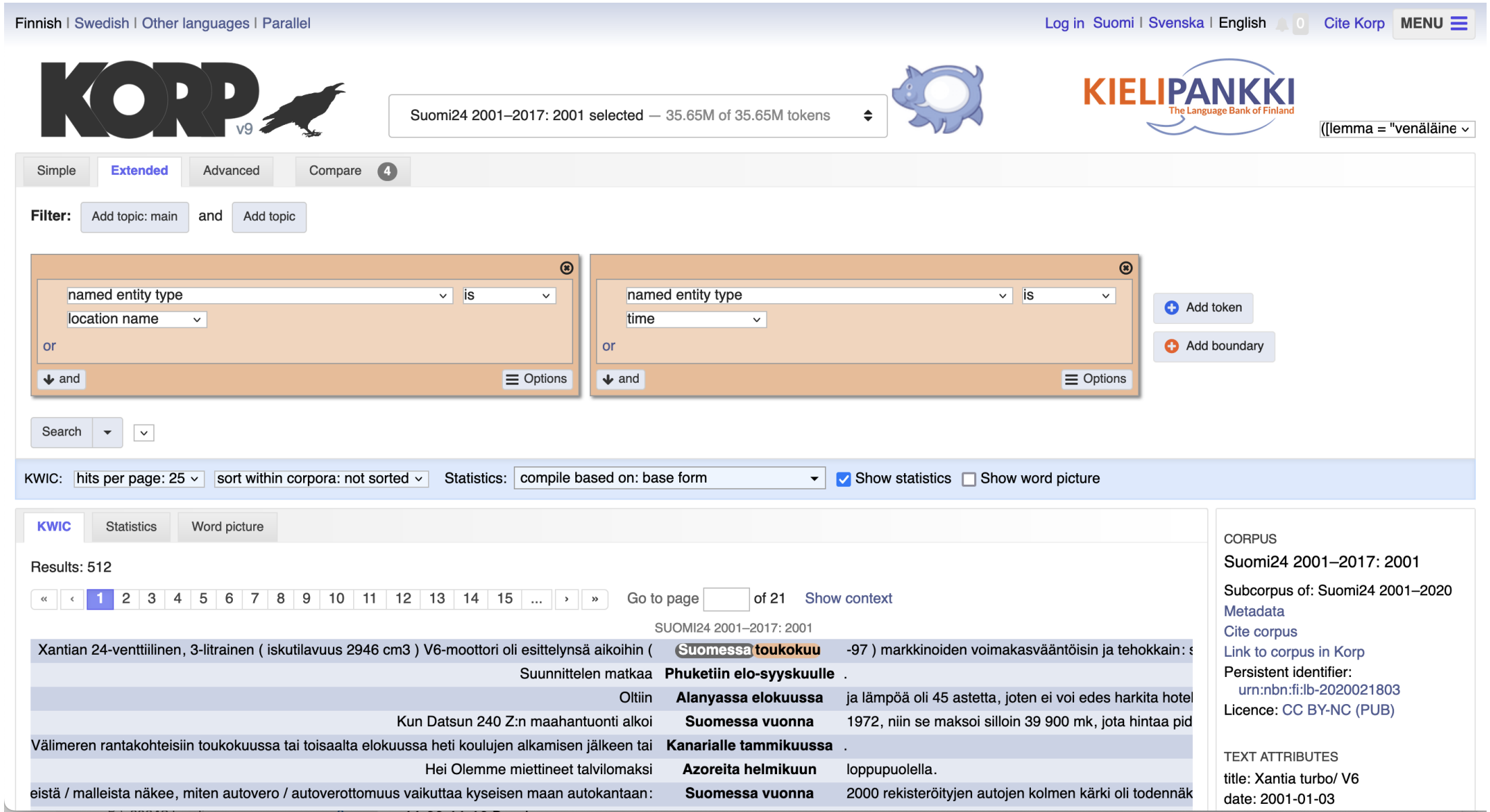


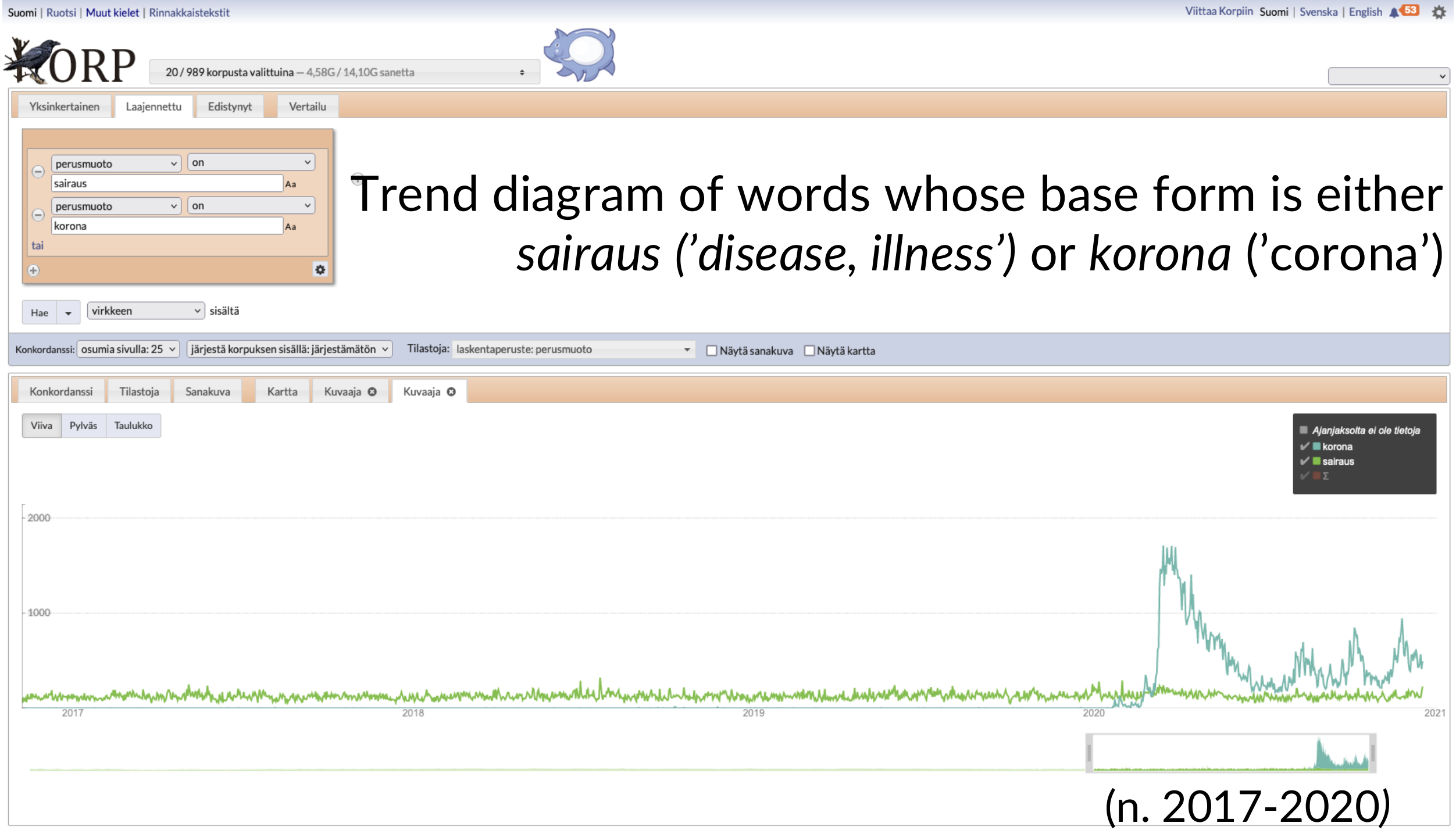
The Suomi 24 Sentences Corpus 2001-2020, Korp version

https://www.kielipankki.fi/korp/



The Suomi 24 Sentences Corpus 2001-2020, Korp version
Other useful annotations: NER, language id, …





Trend diagram of words whose base form is either *sairaus* (’*disease, illness*’) or *korona* (’corona’)

(n. 2017-2020)

Plenary Sessions of the Parliament of Finland: *maahanmuuttaja* 'immigrant'

Sisäasiainministeriön vuoden 2009INE talousarviossa on tavoitteena että	**maahanmuuttajien**	koko koulutusjärjestelmä vastaa sekä laadullisesti että määrällisesti yksilöllisiä tarpei
Tavoitteena on myöskin että	**maahanmuuttajien**	työttömyys puolittuu vuoteen 2012INE mennessä.
nmuuttoalueille laaditaan 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 koulutuksen.
Sinne tuli sekä paikallisia että näitä	**maahanmuuttajia**	pienen ajan päästä tuli valtavasti kumpienkin vanhempia myös mukaan.
On hieno asia se että	**maahanmuuttajia**	ja turvapaikanhakijoita on otettu Suomeen aiempaa enemmän.
jestötoiminta on erittäin keskeisessä asemassa ja tiedämme myöskin miten tärkeitä	**maahanmuuttajat**	ovat meidän jalkapallollemme jos nyt vain viitataan Saksa otteluun.
yrityksiä jotka esimerkiksi olisivat kuin vuokratyöyrityksiä ja tarjoaisivat ja etsisivät	**maahanmuuttajille**	nimenomaan turvapaikkamenettelystä tulleille ihmisille ehkä sitten tällä tavalla työp
varmistamaan se että suomen kielen opetusta on tarjolla ja että sitä saa silloin kun	**maahanmuuttaja**	on valmis sitoutumaan kielen opiskeluun.
n todella paljon ja samaan aikaan meillä sanotaan että maahan tarvitaan niin paljon	**maahanmuuttajia**	ja niin paljon jotta tämä kansa voi selvitä tulevaisuudessa.
steri Väyrynen nostatti pelkotiloja espoolaisissa todeten että metro tuo kaupunkiin	**maahanmuuttajia**	.
Ensimmäinen koskee	**maahanmuuttajien**	valmistavaa opetusta jota esitetään pidennettäväksi puolesta vuodesta vuoden pitui
Lausunnossa valiokunta teki yksitoista toimenpide ehdotusta joista yksi koski	**maahanmuuttajien**	valmistavan koulutuksen lisäämistä ja suomi tai ruotsi toisena kielenä opetusta.
tässä salissa lausuma jossa eduskunta edellytti hallitusta ryhtymään toimenpiteisiin	**maahanmuuttajien**	valmistavan opetuksen ja kielikoulutuksen laajentamiseksi.

« ‹ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ... › »

Download KWIC as Annotations Excel (XLS)

JSON

publication time: 18:06:38
session duration (ms): 10178760
original video
original transcript
name of speaker: Maria Guzenina-Richardson
parliamentary group: parlgroup_sd
speaker role: *[empty]*
type of speech: *[empty]*
speaker: Maria Guzenina-Richardson /sd
utterance begin time (ms): 600928
utterance duration (ms): 11357
utterance end time (ms): 612285
utterance number: 17661
session duration: 2:49:38.760
utterance begin time: 0:10:00.928
utterance end time: 0:10:12.285
utterance duration: 0:00:11.357

Word attributes

base form: maahanmuuttaja
baseform (compound boundaries): maahanmuuttaja
part of speech: noun
msd: NUM_Pl|CASE_Par
dependency relation: direct object
word position in a name: outside (O)

Show Dependency Tree

Show video
Show video in LAT service

Click on link to show video!

Plenary Sessions of the Parliament of Finland, Korp version 1.5: Link to video

Play/Pause | -30 s | -5 s | +5 s | +30 s

Play selection Stop video at selection end:

Media source:
https://down0-80133-down0.dna.qbrick.com/80133-down0/Syysistuntokausi_2008/97._torstaina_23._lokakuuta_2008.mp4

Corpus

Eduskunnan täysistunnot

Persistent Identifier (metadata):
urn:nbn:fi:lb-2019101621

Licence (audio and video): CLARIN PUB +BY +PRIV +ND +OTHER 1.0
Licence (text): CLARIN PUB +BY +PRIV 1.0

Link to search in Korp

Text attributes

publication date: 2008-10-23
publication time: 18:06:38

Please note that the video feature on this page is not supported by all browsers. The page is most likely to work in a Firefox browser.

The original transcript has been aligned with the media file and annotated by using automatic methods. The alignment or the annotations may thus contain errors.

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

name of speaker: Maria Guzenina-Richardson
parliamentary group: parlgroup_sd
speaker role: speakerrole_
type of speech: speechtype_

utterance begin time: 0:10:00.928
utterance end time: 0:10:12.285
utterance duration: 0:00:11.357
utterance begin time (ms): 600928
utterance end time (ms): 612285
utterance duration (ms): 11357

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

List of tools: *www.kielipankki.fi/tools*

Start	Name (and metadata)	Description	Instructions	Install	Info	Administrator	Service level
KORP	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
Aalto-ASR	Aalto University Automatic Speech Recognition System	An automatic speech recognition toolkit that can be used in the CSC computing environment.	Instructions	Install (GitHub)		A?	
ANEE Lexical Networks	**ANEE Lexical Networks**	A graphic semantic dictionary represented as a network. You can use the portal for exploring the meanings of singular Akkadian words in a visual way.					
ANEE Idiolect Network Portal	ANEE Idiolect Network Portal	A portal with over 105,621 pages linked together. The pages contain lists of most similar neighbours, ranked by Double Mutual Rank (DOMUR) similarity measure, for 105,621 cuneiform texts exported from Oracc.					A
Annif	**Annif**	Annif is a tool for automated subject indexing and classification, developed at the National Library of Finland.		Install (GitHub)		1640 KANSALLISKIRJASTO NATIONALBIBLIOTEKET	
	CLARIN Federated Content Search	Run a centralized query from all the resources provided by CLARIN centers.					
COMEDI	**COMEDI**	COMEDI is a Web-based editor for CMDI-conformant metadata, as adopted by CLARIN, hosted by the CLARINO Bergen CLARIN Centre				CLARINO	

KIELIPANKKI
The Language Bank of Finland

Managing access to resources

FIN-CLARIN

CLARIN license categories

Publicly available

Available for academic, logged in users

Access is based on an individual application

Language Bank Rights: lbr.csc.fi

More detailed license conditions

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

+ other resource-specific conditions, if required; e.g., data protection terms and conditions

KIELIPANKKI
The Language Bank of Finland

Language Bank Rights

https://lbr.csc.fi

FIN-CLARIN

What kind of data can be deposited with the Language Bank of Finland?

- Text or speech in any natural language
- Make sure you have the rights to distribute the data, at least for research purposes
  - Copyright, database rights, other Intellectual Property Rights
  - Personal data
- Follow good practices in research ethics
  - When required, ethical review process must be completed prior to data collection

Are safety measures required for sharing?

- The Language Bank of Finland offers various ways for protecting personal data and/or other types of content for restricted use:
  1. **Access management, if needed: university login (ACA); access granted on an individual basis upon application (RES)**
  2. **Resource-specific data protection terms and conditions (must be accepted by the end-users)**
  3. **Data encryption for individuals (example resource: findarc)**
  4. **Sensitive Data (SD) services at CSC**

Suggest a dataset to the Language Bank of Finland!

http://urn.fi/urn:nbn:fi:lb-2021121422

With this form you can ask FIN-CLARIN to publish on-line the essential metadata of the corpus or tool that you wish to deposit with Kielipankki (The Language Bank of Finland) for distribution. The corpus or tool can be completed or still in progress.

- 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 further 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

(*) Name of the information provider *

(*) Email address of the information provider *

ORCID identifier of the information provider (instructions)

What is ORCID?

https://www.clarin.eu/content/depositing-services

Depositing Services

Table of Contents

- Centres Offering Depositing Services
- Other Organisations

One of the fundamental services of the CLARIN infrastructure is making sure that language resources can be archived and made available to the community in a reliable manner. To help researchers to store their resources (e.g. corpora, lexica, audio and video recordings, annotations, grammars, etc.) in a sustainable way, many of the CLARIN centres offer a **depositing service**. They are willing to store the resources in their repository and assist with the **technical and organisational details**. This has a wide range of advantages:

- **Long-term archiving**: a storage guarantee can be given for a long period (up to 50 years in some cases)
- Resources can be cited easily with a **persistent identifier**
- The resources and their metadata will be integrated into the infrastructure, making it possible to **search** for them efficiently
- Password-protected resources can be made available via **an institutional login**

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

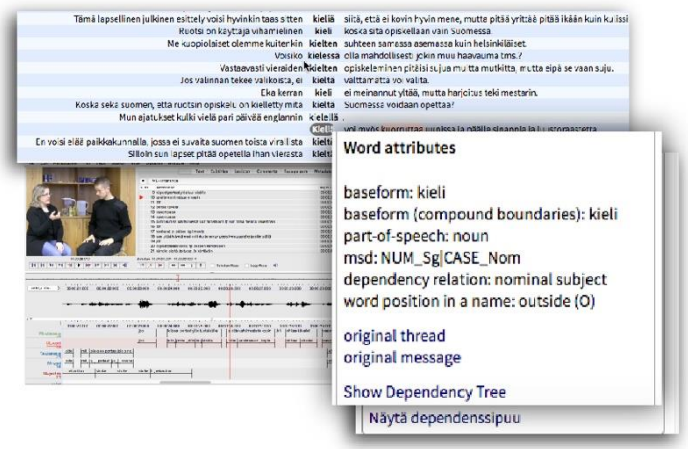
Deposition agreements

Long-term archiving (?)

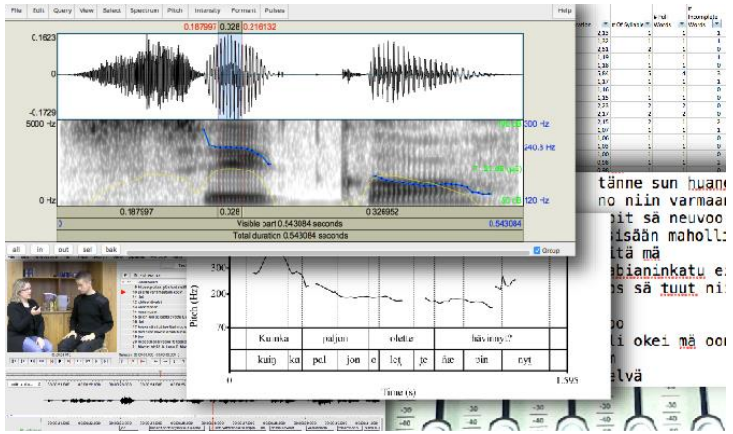
Support for versions and variants

Online courses

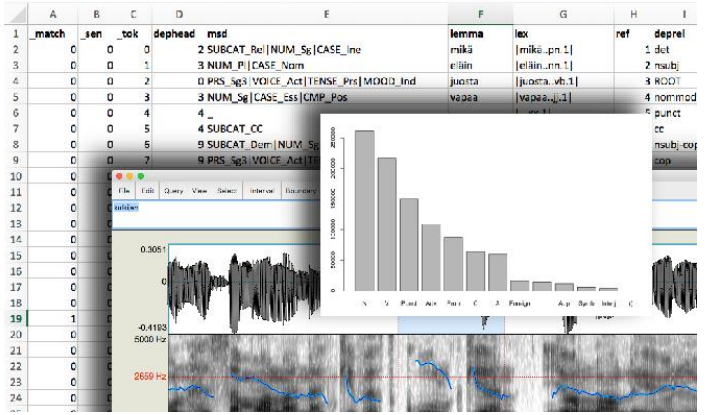
Corpus Linguistics and Statistical Methods (5 cr)



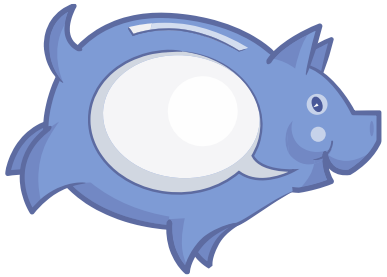
Introduction to Speech Analysis (5 cr)



Data Clinic (5 cr)



The courses are open to all students and researchers within and outside the University of Helsinki, even abroad.



KIELIPANKKI
The Language Bank of Finland

Thank you!

Subscribe to our newsletter:

www.kielipankki.fi

General support
fin-clarin@helsinki.fi

Technical support
kielipankki@csc.fi

Follow us on LinkedIn, Mastodon and YouTube