

**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](mailto:mietta.lennes@helsinki.fi)*



This document is licensed under the  
Creative Commons Attribution 4.0 International license.

**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

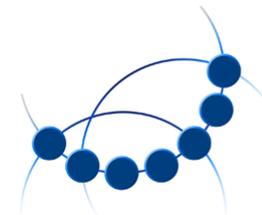
- Researcher of the Month: Heini Kallio (12.3.2026)

# 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.



**CLARIN**  
**CENTRE B**



## 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)

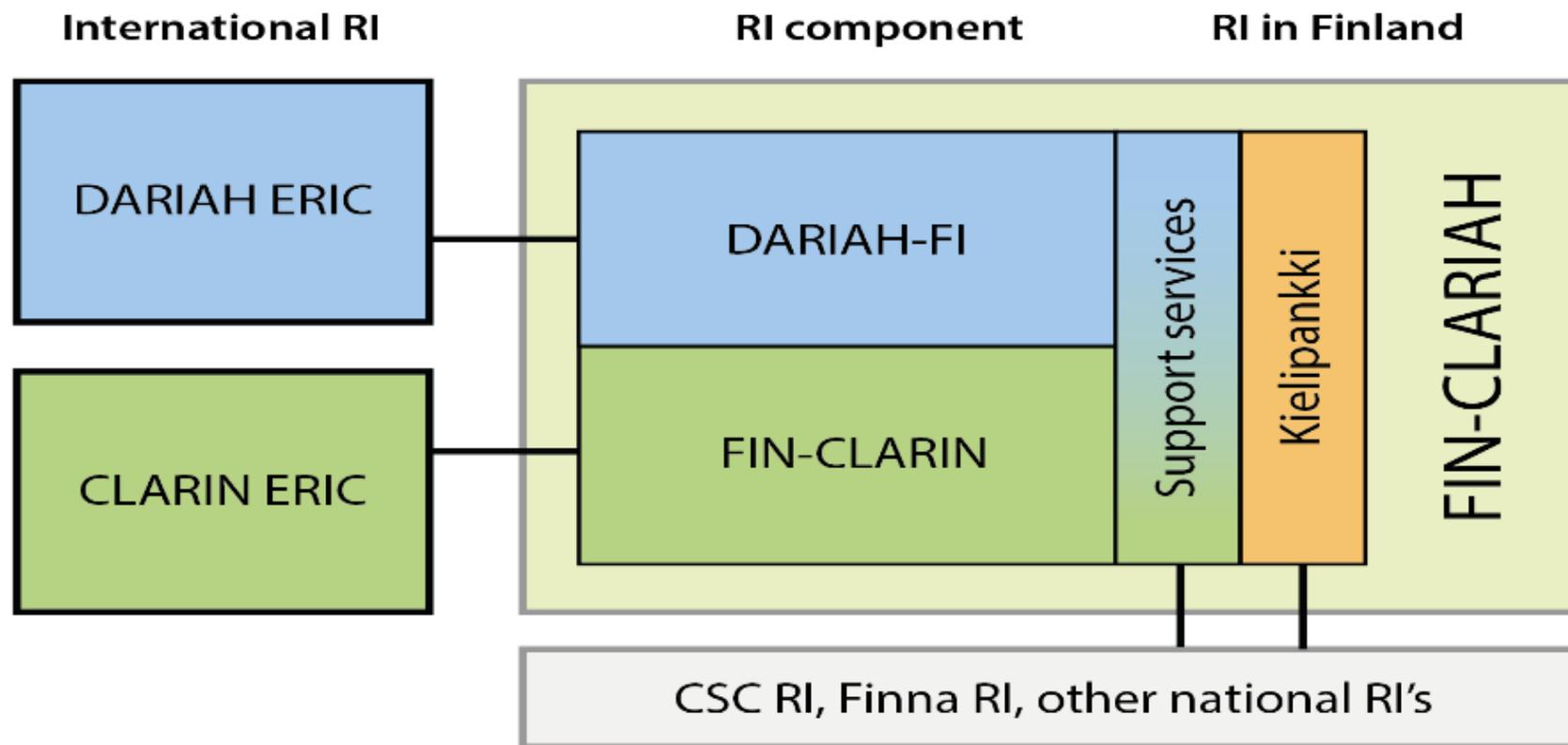
**Findable**

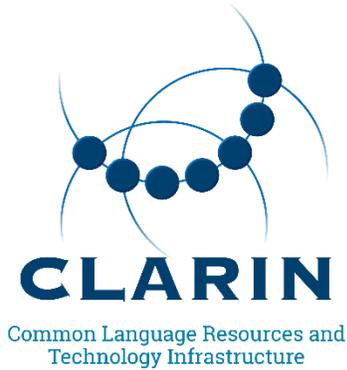
**Accessible**

**FAIR data**

**Interoperable**

**Re-usable**





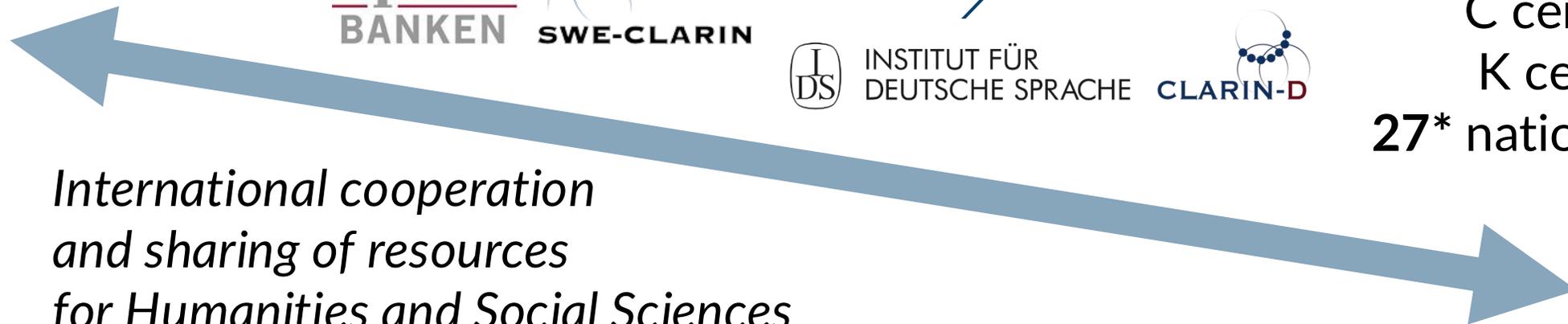
# CLARIN ERIC

European Research Infrastructure Consortium  
founded on February 29, 2012

<https://www.clarin.eu>



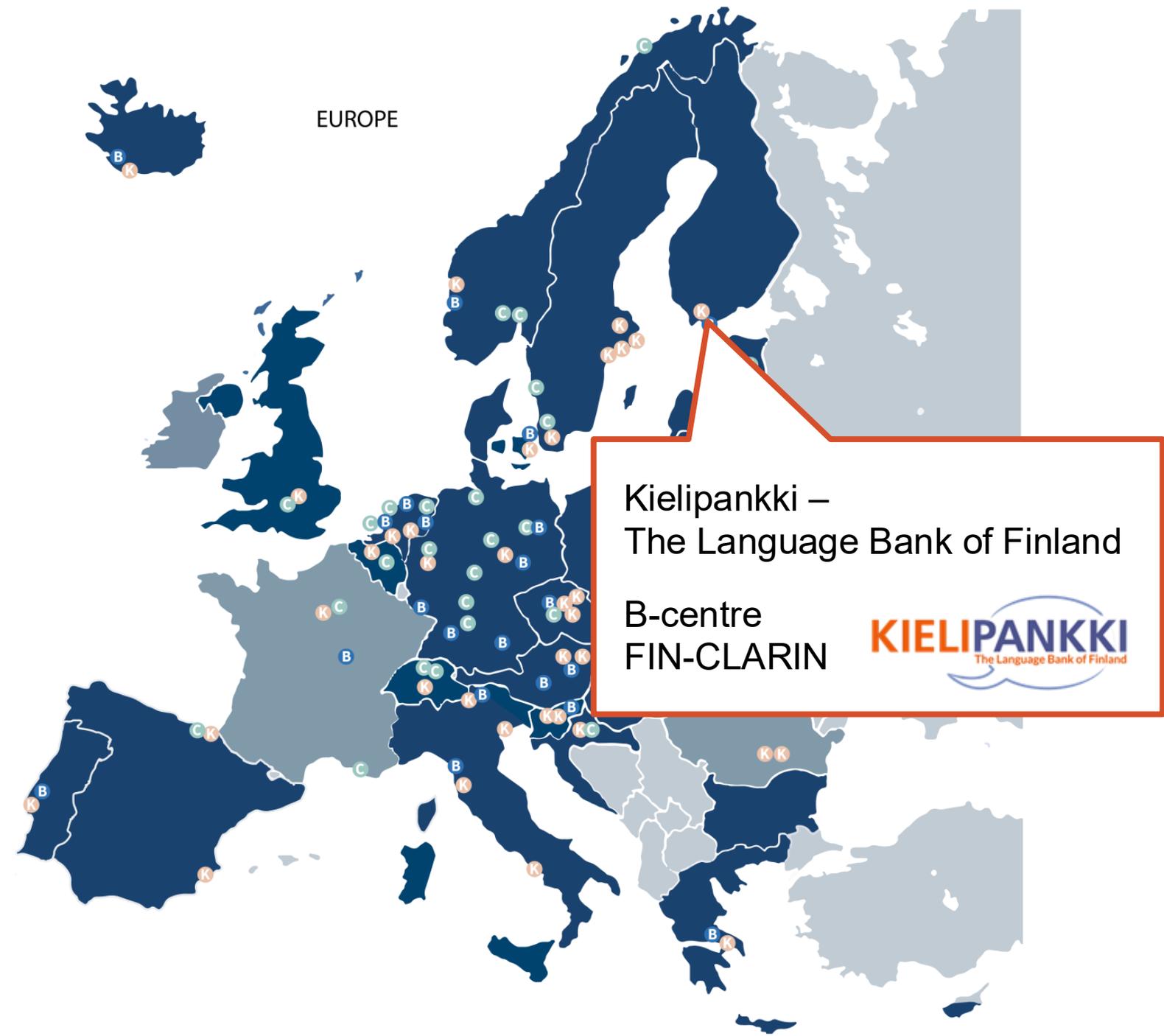
Other  
B centres,  
C centres and  
K centres at  
**27\*** national consortia



*International cooperation  
and sharing of resources  
for Humanities and Social Sciences*



- 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

# <https://csc.fi/>



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)

(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](#)

176 1



## EXMARaLDA Demo corpus

(Part of [Hamburger Zentrum für Sprachkorpora \(HZSK\)](#))

⊕ A selection of short audio and video recordings in various languages



<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](#)

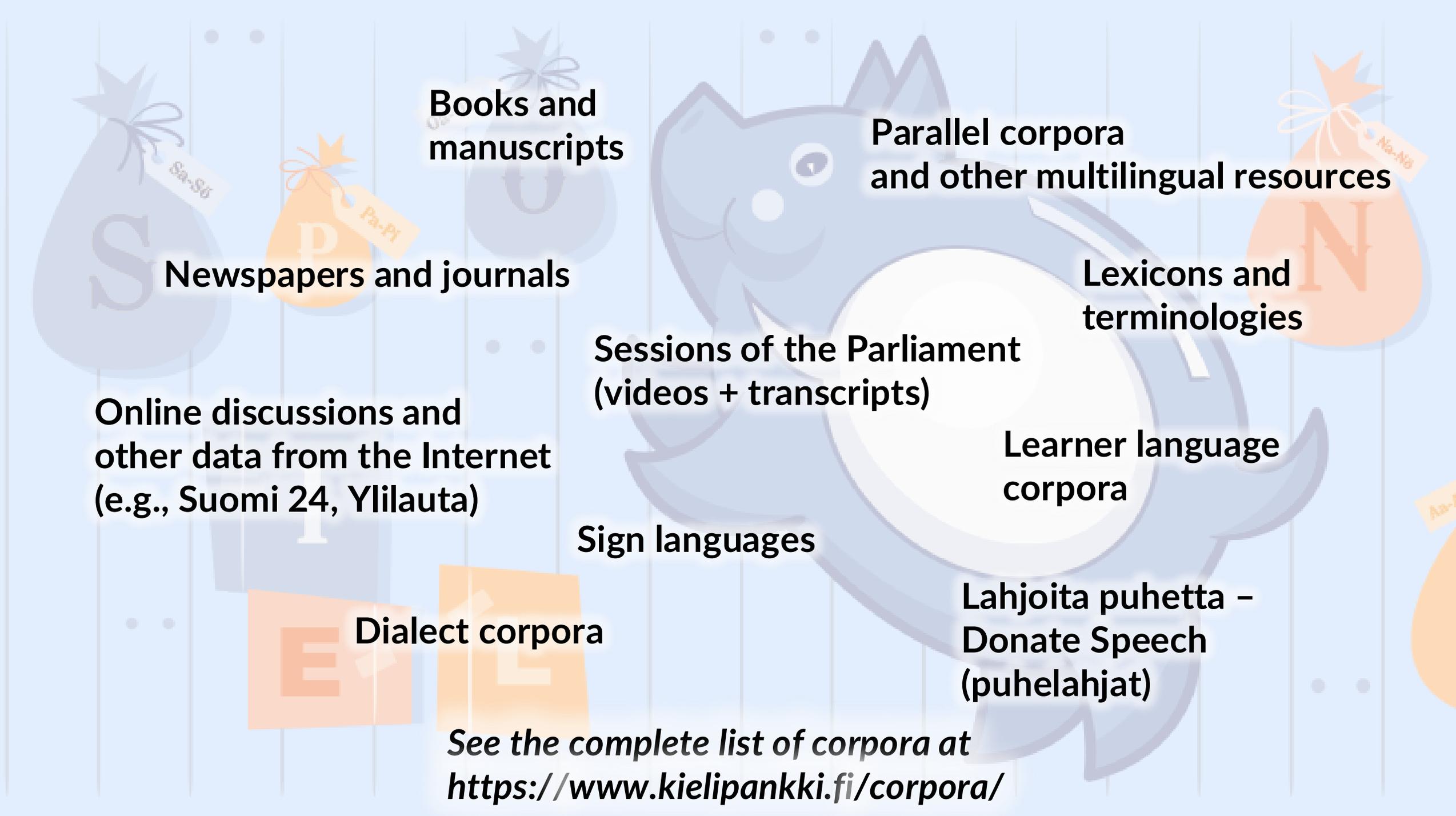
[Tools for normalization](#)

[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**

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

**Sessions of the Parliament  
(videos + transcripts)**

**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 		Korp	”	achemenet		2025	B
agricola-v1-1-korp	The Morpho-Syntactic Database of Mikael Agricola's Works version 1.1, Korp 		Korp	”	agricola		2020	B
ai2d-rst-v1-1	AI2D-RST: A multimodal corpus of 1000 primary school science diagrams version 1.1 		Download	”	ai2d-rst		2020	B
aku-egg-dl	Speech and EGG (Electroglottography) Simultaneous Recordings, downloadable version 		Download	”	aku-egg		2014	B
amph	amph-Corpus 		Download Puhti	”	amph		2008	B
ArkiSyn-korp	ArkiSyn Database of Finnish Conversational Discourse, Helsinki Korp Version 		Korp	”	arkisyn		2017	B
AVOID	Corpus of Age-related Voice Disguise (AVOID)  		Download	”	avoid		2019	B
balt-2025-02-korp	BALT: Babylonian Administrative and Legal Texts - Kielipankki version 2025-02, Korp 		Korp	”	balt		2025	B
BeserCorp	The Corpus of Beserman Udmurt		Korp	”	besercorp		2016	B

# Citation instructions

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

## 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](#)



# Accessing resources



# The Suomi24 Sentences Corpus 2021-2023, Korp version

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

Finnish | Swedish | Other languages | Parallel Log in Suomi | Svenska | English 6 Cite Korp MENU

## KORP v9

23 of 1255 corpora selected — 5.06G of 37.90G tokens

**Simple** Extended Advanced Compare

Filter: Add topic: main and Add topic

ukrainalainen Search

in order and also as  initial part  compound\_mid

KWIC: hits per page: 25 sort within corpora: not sorted

**KWIC** Statistics Word picture

Results: 3,910

« < 1 2 3 4 5 6 7 8 9 10 11 »

**Tietokonevälikkeistä viestintää (36)**

- Suomi24 2001–2023 (23)
  - Suomi24 2001–2017: 2001
  - Suomi24 2001–2017: 2002
  - Suomi24 2001–2017: 2003
  - Suomi24 2001–2017: 2004
  - Suomi24 2001–2017: 2005
  - Suomi24 2001–2017: 2006
  - Suomi24 2001–2017: 2007
  - Suomi24 2001–2017: 2008
  - Suomi24 2001–2017: 2009
  - Suomi24 2001–2017: 2010
  - Suomi24 2001–2017: 2011
  - Suomi24 2001–2017: 2012
  - Suomi24 2001–2017: 2013
  - Suomi24 2001–2017: 2014
  - Suomi24 2001–2017: 2015
  - Suomi24 2001–2017: 2016
  - Suomi24 2001–2017: 2017
  - Suomi24 2018–2020: 2018

**Suomi24 2001–2023**

Suomi24 virkkeet -korpus 2001–2023, Korp-versio

Suomi24-keskustelupalvelun keskustelut vuosilta 2001–2023 (1.1.2001–31.12.2023). Aineistossa näkyy kaikkien keskustelujen sisältö enintään kappaletasolla. Aineisto on jaettu osakorpuksiin vuosittain. Tutkijat voivat myös ladata käyttöönsä koko Suomi24 2001–2023 -aineiston Kielipankin latauspalvelusta (lisenssi).

Tämä kokoelma sisältää seuraavat aineistot:

- [Suomi24 virkkeet -korpus 2001–2017, Korp-versio 1.3](#)
- [Suomi24 virkkeet -korpus 2018–2020, Korp-versio 1.1](#)
- [Suomi24 virkkeet -korpus 2021–2023, Korp-versio](#)

2025-04-07: Kokoelmaan on lisätty vuosien 2021–2023 keskustelut ([Suomi24 virkkeet -korpus 2021–2023, Korp-versio](#)). Lisäksi koko aineistoon on merkitty **FINER 1.6 -nimientunnistimella** tunnistetut nimet ja **HeLI-OTS 2.0 -kielentunnistimella** tunnistetut virkkeiden kielet. Vuosien 2001–2020 aineistoissa on myös poistettu ylimääräisiä välilyöntejä aihealueista, otsikoista ja kirjoittajien nimimerkeistä. Näiden lisäysten ja muutosten vuoksi vanhempien osien versionumeroita on kasvatettu. [Tarkempia tietoja muutoksista.](#)

ukrainalainen

–2017: 2002  
Suomi24-2001-2017-  
Suomi24 2001–2023

page  
Korp  
fier:  
020021803  
-NC 4.0 (PUB)

ES  
mysteeri  
1

# The Suomi24 Sentences Corpus 2021-2023, Korp version

## <https://www.kielipankki.fi/korp/>

Finnish | Swedish | Other languages | Parallel

Log in Suomi | Svenska | English  6 Cite Korp MENU 



23 of 1255 corpora selected — 5.06G of 37.90G tokens



ukrainalainen 

Simple Extended Advanced Compare

Filter: Add topic: main and Add topic

ukrainalainen|

Search 

ukrainalainen (noun)

**ukrainalainen (adjective) 31561**

ukrainalaisenositas (noun)

ukrainalainen nainen (noun)

ukrainalainenfasisti (noun)

ukrainalainenmies (noun)

ukrainalainenjoukko (noun)

ukrainalaisensyntyinen (adjective)

ukrainalainen (adverb)

ukrainalainenhävittäjä (noun)

ound\_middle  final part and  case-insensitive

Sort: not sorted 

Step

compile based on: word 

Show statistics

Show word picture

10 11 12 13 14 15 ...   Go to page  of 157 Show context

SUOMI24 2001–2017: 2002

Wolf Messing oli syntyjään **ukrainalainen** köyhän kodin poika, jonka isä halusi poikansa menevän luostarik

Keskuskomitean 62 jäsentä olivat viisi venäläistä, yksi **ukrainalainen**, kuusi latvialaista, kaksi saksalaista, yksi tsekki, kaksi armenialai

Wolf Messing oli syntyjään **ukrainalainen** köyhän kodin poika, jonka isä halusi poikansa menevän luostarik

Wolf Messing oli syntyjään **ukrainalainen** köyhän kodin poika, jonka isä halusi poikansa menevän luostarik

hyvinä esimerkkeinä ovat keihäänheittäjä Harri Haatainen ja **ukrainalainen** moninkertainen maailmanmestari kuulantyyntönnössä ( nimi ei nyt tu

Ernud ( Enroth? ) livari Aleksejevits, s. 1902, Viipurin lääni, Suomi; **ukrainalainen** ( ? ); rappaaja, Kontupohjan paperikombinaatti;

Ernud ( Enroth? ) livari Aleksejevits, s. 1902, Viipurin lääni, Suomi; **ukrainalainen** ( ? ); rappaaja, Kontupohjan paperikombinaatti;

SUOMI24 2001–2017: 2003

itkovskaja on venäläinen, mutta olen **saanut** selville, että todellisuudessa hän on kansallisuudeltaan **ukrainalainen** ( ihan eri asia kuin venäläinen ) ja syntynyt New Yorkissa.

Vai että **ukrainalainen** ..

Ei tarvitse olla mikään Saimi Nousiais tyylinen kukkakeppi, tai mikään **ukrainalainen** kuulantyyntöntäjä, vaan juuri kivasti siltä väliltä, kyllä sen ihanneko

CORPUS

Suomi24 2001–2017: 2003

Short name: suomi24-2001-2017-korp-v1-3

Subcorpus of: Suomi24 2001–2023

Metadata

[Cite corpus](#)

[Resource group page](#)

[Link to corpus in Korp](#)

Persistent identifier:  
urn:nbn:fi:lb-2020021803

Licence: CC BY-NC 4.0 (PUB)

TEXT ATTRIBUTES

title: Tsetseenityön murhaaja Juri Budanov vapautettu

date: 2003-01-01

time: 17:31:19

S

# The Suomi 24 Sentences Corpus 2001-2020, Korp version

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

## Extended search

Finnish | Swedish | Other languages | Parallel Log in Suomi | Svenska | English 11 Cite Korp MENU



20 of 1246 corpora selected — 4.58G of 37.30G tokens



ukrainalainen

Simple **Extended** Advanced Compare 4

Filter: Add topic: main and Add topic

base form is

Venäjä Aa

or

case-sensitive  
case-insensitive

↓ and Options

word is

<any word> Aa

or

Repeat 0 to 20 times Options

base form is

Ukraina Aa

or

↓ and Options

Search within sentence

KWIC: hits per page: 25 sort within corpora: not sorted

Statistics: compile based on: base form  Show statistics  Show word picture

# The Suomi 24 Sentences Corpus 2001-2020, Korp version

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

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

KWIC

Statistics

Word picture

Graph 

Results: 26,205

« < 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ... > » Go to page  of 1049 Show context

SUOMI24 2001–2017: 2001

nettisivuillaan ennusti mm. että vuoden 2000 aikana **Venäjä, Valko-Venäjä ja Ukraina** solmivat valtioliiton, Suomessa alkaa "hysterinen l

SUOMI24 2001–2017: 2002

**Venäjällä otettiin joulukuussa käyttöön 1000 MW:n Rostov-ydinvoimalayksikkö lähellä Ukrainan** rajaa.

Nykyisellä **Venäjällä, Valko-Venäjällä ja Ukrainassa** asuu yhteensä 5 miljoonaa ihmistä alueilla, joilla alk

den luokitus ( cesium-137, kBq/m2 ) Alueiden laajuus **Venäjällä, Valko- venäjällä ja Ukrainassa** yhteensä ( km2 ) Asukasmäärä 1995 Alueiden kuva

Uusnoitus ja taikuus kasvoi Virossa, **Venäjällä ja Ukrainassa**, kun talous ja turvallisuus menivät.

Matkailua välillä **Venäjä - Ukraina** - Romani- Jugoslavia - Makedonia jne.

ps. helpotusta viljapulaan tulee lähivuosina hieman **Venäjän ja Ukrainan** mustan mullan alueelta ( arvioitu potentiaali 40Milj.t

armastikin ostaa / varastaa valmis ydinkärki vaikkapa **Venäjän tai Ukrainan** varastoista.

Kokouksessa oli osallistujia Suomen ja **Venäjän lisäksi mm. Ranskasta, Saksasta, Italiasta, Sloveniasta, Romaniasta, Moldovasta, Ukrainasta**, Georgiasta, Armeniasta ja Japanista.

Ongelma on vaan se, että **Venäjällä ja Ukrainassa** on aivan älyttömän paljon juutalaisia.

että juutalaisten prosenttiosuus Euroopan puoleisella **Venäjällä, Valko-Venäjällä ja Ukrainassa** on ollut likemmäs 20 %.

Jos kuvittelet, että **Venäjän ludmila tai Ukrainan** olia katsoo hyvällä ryppyturneita kavereitten kanssa

SUOMI24 2001–2017: 2003

Vähemmän kuin **Venäjällä tai Ukrainassa** .

Onhan kuitenkin totta, että kommunismi **Venäjällä ja Ukrainassa** on nyt jo onneksi kaatunut ja menneisyyttä.

n pääministeri Silvio Berlusconi ehdotti maaliskuussa **Venäjän, Israelin, Ukrainan**, Moldovan, ja Turkin hyväksymistä EU:n jäseniksi.

si neuvosto on hyväksynyt kolme yhteistä strategiaa: **Venäjä, Ukrainaa** ja Välimeren aluetta koskevat yhteiset strategiat.

Vähemmän kuin **Venäjällä tai Ukrainassa** .

ävä eron jakolinjoistaan ja laajennuttava esimerkiksi **Venäjälle, jolla on sotilaallista voimaa sekä Ukraina**, Moldovaan, Turkkiin ja jopa Israeliin ", Berlusconi v

# The Suomi 24 Sentences Corpus 2001-2020, Korp version

## Other useful annotations: NER, language id, ...

Finnish | Swedish | Other languages | Parallel Log in Suomi | Svenska | English 0 Cite Korp MENU

**KORP** v9  Suomi24 2001–2017: 2001 selected — 35.65M of 35.65M tokens  **KIELIPANKKI** The Language Bank of Finland (lemma = "venäläine" v)

Simple **Extended** Advanced Compare 4

Filter: Add topic: main and Add topic

named entity type is location name

or

and Options

named entity type is time

or

and Options

+ Add token

+ Add boundary

Search [ ] [ ]

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

**KWIC** Statistics Word picture

Results: 512

« < 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ... > » Go to page [ ] of 21 Show context

SUOMI24 2001–2017: 2001

Xantian 24-venttiilinen, 3-litrainen ( iskutilavuus 2946 cm3 ) V6-moottori oli esittelynsä aikoihin ( Suomessa toukokuu -97 ) markkinoiden voimakasvääntöisin ja tehokkain: s	
Suunnittelen matkaa Phuketiin elo-syyskuulle .	
Oltiin Alanyassa elokuussa ja lämpöä oli 45 astetta, joten ei voi edes harkita hotel	
Kun Datsun 240 Z:n maahantuonti alkoi Suomessa vuonna 1972, niin se maksoi silloin 39 900 mk, jota hintaa pid	
Välimeren rantakohteisiin toukokuussa tai toisaalta elokuussa heti koulujen alkamisen jälkeen tai Kanarially tammikuussa .	
Hei Olemme miettineet talvilomaksi Azoreita helmikuun loppupuolella.	
eistä / malleista näkee, miten autovero / autoverottomuus vaikuttaa kyseisen maan autokantaan: Suomessa vuonna 2000 rekisteröityjen autojen kolmen kärki oli todennäk	

CORPUS

Suomi24 2001–2017: 2001

Subcorpus of: Suomi24 2001–2020

Metadata

Cite corpus

Link to corpus in Korp

Persistent identifier:  
urn:nbn:fi:lb-2020021803

Licence: CC BY-NC (PUB)

TEXT ATTRIBUTES

title: Xantia turbo/ V6

date: 2001-01-03



20 / 989 korpusta valittuina — 4,58G / 14,10G sanetta



Yksinkertainen Laajennettu Edistynyt Vertailu

perusmuoto on  
sairaus Aa  
perusmuoto on  
korona Aa  
tai

Hae virkkeen sisältä

Konkordanssi: osumia sivulla: 25 järjestä korpuksen sisällä: järjestämätön Tilastoja: laskentaperuste: perusmuoto  Näytä sanakuva  Näytä kartta

Konkordanssi Tilastoja Sanakuva Kartta Kuvaaja Kuvaaja

Viiva Pylväs Taulukko



(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 tarpeita. 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. **maahanmuuttajat** ovat meidän jalkapallollemme jos nyt vain viitataan Saksa otteluun. **maahanmuuttajille** nimenomaan turvapaikkamenettelystä tullee ihmisille ehkä sitten tällä tavalla työp **maahanmuuttajia** on valmis sitoutumaan kielen opiskeluun. **maahanmuuttajia** ja niin paljon jotta tämä kansa voi selvitä tulevaisuudessa. **maahanmuuttajia** . **maahanmuuttajien** valmistavaa opetusta jota esitetään pidennettäväksi puolesta vuodesta vuoden pitui **maahanmuuttajien** valmistavan koulutuksen lisäämistä ja suomi tai ruotsi toisena kielenä opetusta. **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

JSON

publication time: 10.00.00  
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\_PI|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:

**Utterance:** Sitten äskettäin Espoossa pidetyssä vaalitulaisuudessa 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

## Media source:

[https://down0-80133-down0.dna.qbrick.com/80133-down0/Syysistuntokausi\\_2008/97.\\_torstaina\\_23.\\_lokakuuta\\_2008.mp4](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.*

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

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

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

Start	Name (and metadata)	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.	<a href="#">Instructions</a>		<a href="#">?</a>		A
Download	Download service	Download certain corpora.			<a href="#">?</a>		A
Aalto-ASR	<a href="#">Aalto University Automatic Speech Recognition System</a>	An automatic speech recognition toolkit that can be used in the CSC computing environment.	<a href="#">Instructions</a>	<a href="#">Install (GitHub)</a>	<a href="#">?</a>	<b>A?</b>	
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.			<a href="#">?</a>		
ANEE Idiolect Network Portal	<a href="#">ANEE Idiolect Network Portal</a>	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 href="#">?</a>		A
Annif	Annif	Annif is a tool for automated subject indexing and classification, developed at the National Library of Finland.		<a href="#">Install (GitHub)</a>	<a href="#">?</a>		
	CLARIN Federated Content Search	Run a centralized query from all the resources provided by CLARIN centers.			<a href="#">?</a>		
COMEDI	COMEDI	COMEDI is a Web-based editor for CMDI-conformant metadata, as adopted by CLARIN, hosted by the CLARINO Bergen CLARIN Centre			<a href="#">?</a>		



# Managing access to resources



# 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



# Language Bank Rights

*<https://lbr.csc.fi>*



# 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?](#)

# Depositing Services



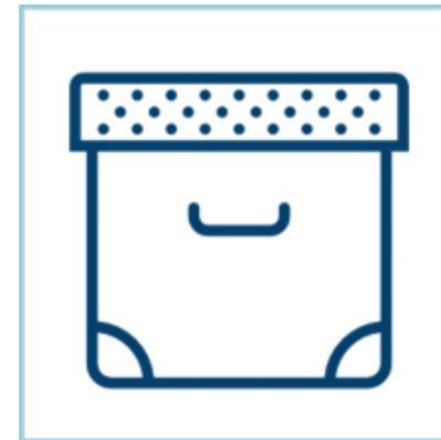
## Table of Contents

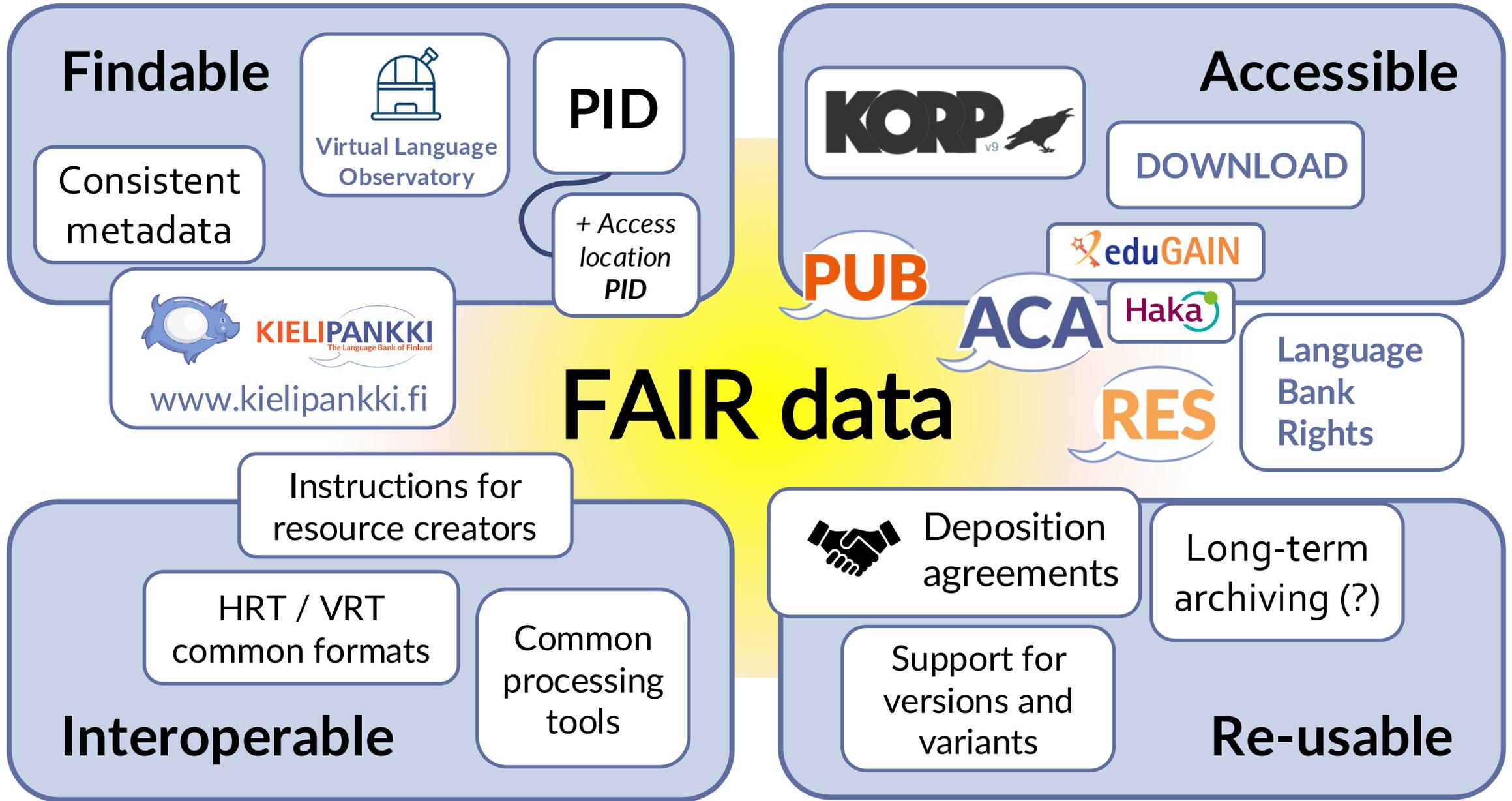
Centres Offering Depositing Services

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**





## Findable

Consistent metadata



Virtual Language Observatory

PID

+ Access location  
PID



**KIELIPANKKI**  
The Language Bank of Finland

[www.kielipankki.fi](http://www.kielipankki.fi)

## Accessible



DOWNLOAD

PUB

eduGAIN

ACA

Haka

RES

Language Bank Rights

# FAIR data

## Interoperable

Instructions for resource creators

HRT / VRT common formats

Common processing tools



Deposition agreements

Long-term archiving (?)

Support for versions and variants

## Re-usable

## 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)

Visible part 0.543984 seconds  
Total duration 0.543084 seconds

kuinka peli on istä se lähtee nyt



## Data Clinic (5 cr)

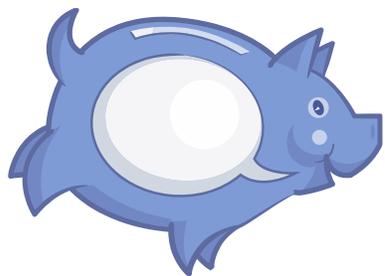
A	B	C	D	E	F	G	H	I
1	_match	_sen	_tokk	dephead	msd	lemma	lex	ref
2	0	0	0	2	SUBCAT_Re NUM_Sg CASE_ine	mikä	[mikä.pn.1]	1
3	0	0	1	3	NUM_P CASE_Nom	eiäin	[eiäin.rm.1]	2
4	0	0	2	0	PRS_Sg3 VOICE_Act TENSE_Prs MOOD_Ind	juosta	[juosta.vb.1]	3
5	0	0	3	3	NUM_Sg CASE_ess CMP_Pos	vapaa	[vapaa.jt.1]	4
6	0	0	4	4				5
7	0	0	5	4	SUBCAT_CC			6
8	0	0	5	9	SUBCAT_Dum NUM_Sg			7
9	0	0	7	9	PRS_Sg3 VOICE_Act T			8
10								9
11								10
12								11
13								12
14								13
15								14
16								15
17								16
18								17
19	1							18
20	0							19
21	0							20
22	0							21
23	0							22
24	0							23



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



DIGITAL HUMANITIES  
COURSE REGISTRY



**KIELIPANKKI**  
The Language Bank of Finland

Thank you!

Subscribe to our newsletter:

[www.kielipankki.fi](http://www.kielipankki.fi)

**General support**

*fin-clarin@helsinki.fi*

**Technical support**

*kielipankki@csc.fi*

Follow us on LinkedIn, Mastodon and YouTube

