

KIELIPANKKI
The Language Bank of Finland

Kielipankki – The Language Bank of Finland

Resources and services provided by FIN-CLARIN

Mietta Lennes, University of Helsinki



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.



Researcher of the Month: Eero Voutilainen

Organization



Who are the Language Bank?

Support



Help and instructions.

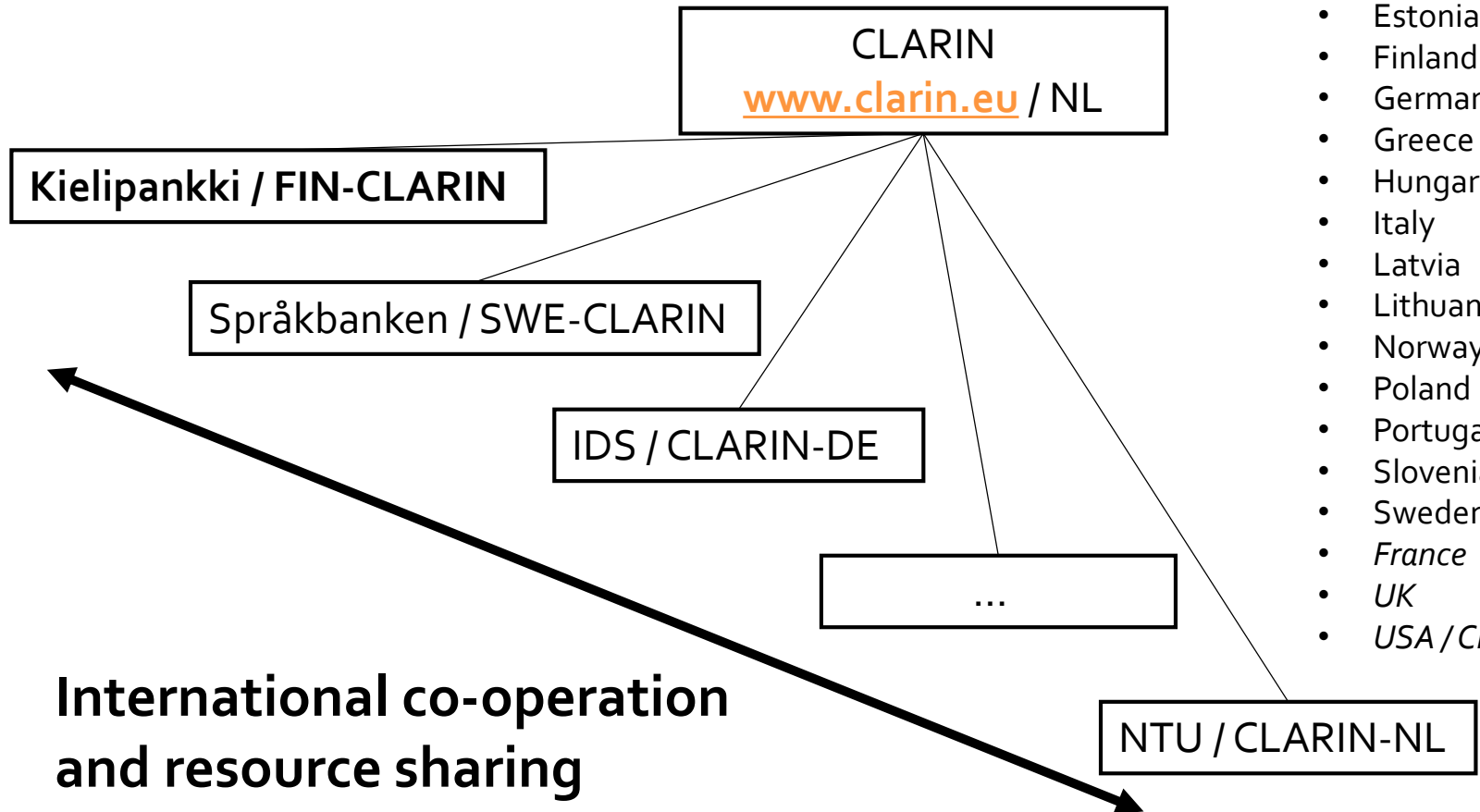
News

- Researcher of the Month: Eero Voutilainen (10.11.2017)
- FIN-CLARIN is looking for a project planning officer (23.10.2017)
- Researcher of the Month: Markus Juutinen (10.10.2017)
- Network maintenance 10.10.2017

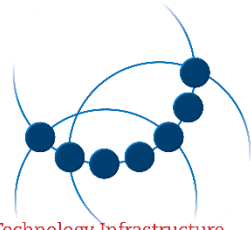


CLARIN ERIC

European Research Infrastructure Consortium
founded 29. February 2012



- **The Netherlands**
- Austria
- Bulgaria
- Czech Republic
- Denmark
- DLU
- Estonia
- Finland
- Germany
- Greece
- Hungary
- Italy
- Latvia
- Lithuania
- Norway
- Poland
- Portugal
- Slovenia
- Sweden
- *France*
- *UK*
- *USA / CMU*



FIN-CLARIN partners

- University of Helsinki
- CSC – IT Center for Science

Coordinates and provides access to centralized corpora 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

Provides access to locally developed corpora and tools



Users of the Language Bank

- Researchers from all fields welcome!
- Many corpora are available even without signing in.
- FIN-CLARIN can help you publish your own corpus.





Corpora

Gw = billion words, Mw = million words, h = hours

Resources	2017	2022
Text		
Magazines and newspapers 1770- (NLF and Web publ.)	12 Gw	20 Gw
Social media and similar sources 2000- (Suomiz4, Ylilauta, ...)	4 Gw	10 Gw
Literature and manuscripts (Gutenberg, Fennica, archives)	60 Mw	70 Mw
Speech		
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
Multilingual and Other Resources		
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



Corpora

Gw = billion words, Mw = million words, h = hours

Resources	2017	2022
Text		
Magazines and newspapers 1770- (NLF and web publ.)	12 Gw	20 Gw
Social media and similar sources 2000- (Twitter, Ylilauta, ...)	4 Gw	10 Gw
Literature and manuscripts (Gutenberg and other archives)	60 Mw	70 Mw
Speech		
News broadcasts (YLE)		10000 h
Video sessions from the Finnish Parliament	500 h	1000 h
Dialect and everyday speech (Kotus, Aalto)	500 h	1000 h
Sign language resources (Aalto, Helsinki)	20 h	500 h
Multilingual and Other Resources		
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

Currently
about 18 GW
in more than
650 databases



Korp: Corpus of Contemporary American English (COCA)

KORP SEARCH: 'TERRORIST'



Korp search interface: Suomi24 corpus



10 of 657 corpora selected — 2.66G of 8.58G tokens

terroristi..nn.1

Simple Extended Advanced Compare

terroristi (noun)

Search

also as initial part final part and case-insensitive

in sentences which contain

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

KWIC Statistics Word picture Map Name classification

Results: 68,735

« < 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ... > » Go to page of 2750

Show context

Suomi24 (1/10)

naapurimaan aluetta tykein ja ohjuksin ja vie jopa aseita naapurin **terroristeille** .
 Alla joku väitti, että muslimeista 3 promillea muslimeista on **terroristeja** .
 Jos siis otetaan kaikki muslimit mukaan **terroristeja** on 5 500 000 *1.1/100 *3/1000= 180. Eikö 180 ter
 t mukaan terroristeja on 5 500 000 *1.1/100 *3/1000= 180. Eikö 180 **terroristia** ole liikaa Suomeen?
 Tottakai JOS Kaikki olisivat **terroristeja** .
 Alla joku väitti, että muslimeista 3 promillea muslimeista on **terroristeja** .
 Jos siis otetaan kaikki muslimit mukaan **terroristeja** on 5 500 000 *1.1/100 *3/1000= 180. Eikö 180 ter
 t mukaan terroristeja on 5 500 000 *1.1/100 *3/1000= 180. Eikö 180 **terroristia** ole liikaa Suomeen?
 Ja sinähän tiedät että muslimeista **terroristeja** on jotain 3 promillea?
 Aka **terroristien** kanssa ei neuvotella.
 Kiovan rikollisen junan **terroristit** aloittivat yöllä 02.00 ensin tulituksen vapausarm
 yöllä 02.00 ensin tulituksen vapausarmeijan aseisiin ja sen jälkeen

Corpus

Suomi24 (1/10)

Metadata

Licence: CC BY-NC (CLARIN PUB)

Cite corpus

Link to corpus in Korp: urn:nbn:fi:lb-2015120401

Text attributes

title: Nato uhittelee Venäjän rajoilla lukuisissa...

title word lemmas: nato uhitella Venäjä raja lukuisa ...

date: 10.06.2015

text_time: 10:09

discussion thread id: 13631623



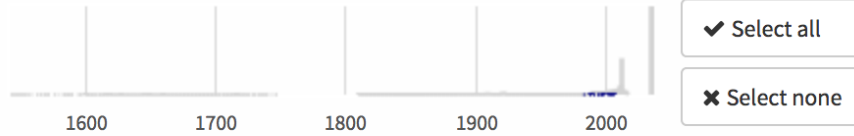
Corpora in other languages



terroristi..nn.1

Simple Extended

MULCOLD englanti selected — 359.87K of 3.45G tokens



- ORACC (6)
- ERME (2)
- Fenno-Ugrica (10)
- English / Englanti (157)
 - E-thesis (25)
 - COCA: Corpus of Contemporary American English (beta) (5)
 - COHA: Corpus of Historical American English (beta) (75)
 - GloWbE: Global Web-based English (beta) (40)
 - ScotsCorr (beta) (9)



MULCOLD englanti

ELFA

Topling (English) [RES]

- Deutsch / Saksa / German (2)
- Français / Ranska / French (1)
- Español / Espanja / Spanish (1)
- Русский / Venäjä / Russian (7)
- Helsinki Corpus of Swahili 2.0 (HCS 2.0) (4)
- SUS-kenttätyö (näyte) (3)

Show map

COCA: Corpus of Contemporary American English (beta)

COCA: Corpus of Contemporary American English – Kielipankki Korp version 2017H1 (beta)

The COCA corpus contains about 520 million words in 220,000 texts of US English from the years 1990–2015. The corpus is evenly divided into spoken, fiction, magazine, newspaper and academic genres.

Note: To follow the US Fair Use Law, every 200 words, ten words have been removed and replaced with “@” ([more information](#)).

Metadata

Licence: [ACA-Fi \(Academic users in Finland\)](#)

[Cite corpus](#)



Log in to use an ACA-licensed corpus

Swedish | Other languages | Parallel Cite Korp Suomi Svenska English 8

KORP

0 of 657 corpora selected — 0 of 8.58G tokens

Basic | Extended | Advanced | Compare

Search

initial part final part and case-insensitive

Sentences which contain

Items per page: 25 sort within corpora: not sorted Statistics: compile based on: word Show word picture Show map

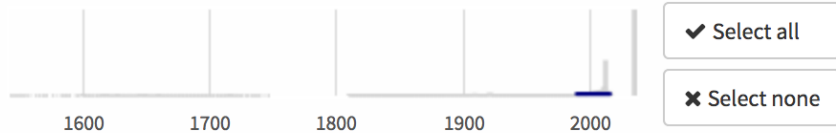
- About Korp
- Help
- Korp Labs (Korp Beta)
- Annotation Lab
- Multilingual Annotation Lab (Sparv)
- Feedback
- Privacy policy
- CLARIN Terms of Service
- The Language Bank of Finland
- Log in



Now select the COCA corpus in the corpus menu



5 of 195 corpora selected — 624.40M of 3.45G tokens



- ORACC (6)
- ERME (2)
- Fenno-Ugrica (10)
- English / Englanti (157)
 - E-thesis (25)
 - COCA: Corpus of Contemporary American English (beta) (5)
 - COHA: Corpus of Historical American English (beta) (75)
 - GloWbE: Global Web-based English (beta) (40)
 - ScotsCorr (beta) (9)
 - MULCOLD englanti
 - ELFA
 - Topling (English) [RES]
- Deutsch / Saksa / German (2)
- Français / Ranska / French (1)
- Español / Espanja / Spanish (1)
- Русский / Venäjä / Russian (7)
- Helsinki Corpus of Swahili 2.0 (HCS 2.0) (4)
- SUS-kenttätyö (näyte) (3)

terroristi..nn.1

✓ Select all

✗ Select none

Show map

COCA: Corpus of Contemporary American English (beta)

COCA: Corpus of Contemporary American English – Kielipankki Korp version 2017H1 (beta)

The COCA corpus contains about 520 million words in 220,000 texts of US English from the years 1990–2015. The corpus is evenly divided into spoken, fiction, magazine, newspaper and academic genres.

Note: To follow the US Fair Use Law, every 200 words, ten words have been removed and replaced with “@” ([more information](#)).

Metadata

Licence: ACA-Fi (Academic users in Finland)

Cite corpus



Search for any form of the word 'terrorist' ("Extended" tab)

Finnish | Swedish | Other languages | Parallel



2 of 66 corpora selected — 243.45M of 494.58M tokens

Simple

Extended

Advanced

Compare

baseform

is

terrorist

Aa

or

case-insensitive

Search

within

sentence

KWIC: hits per page: 25

sort within corpora: not sorted

Statistics: compile based on: word

Show

KWIC

Statistics

Map

Name classification

Graph



'terrorist'

baseform is
 Aa
 or

Search within

KWIC: Statistics: Show map

KWIC Statistics Map Name classification Graph

Results: 5,392

« < 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ... > » Go to page of 216 Show context

COCA: ACADEMIC (BETA)

carrying the ANC delegation swept by (the irony of the police escort for these former " **terrorists** " was not missed), as both parties expressed a sincere desire to move away from
 ect the rights of @ @ @ @ @ @ @ @ @ @ kind of regional, racial and religious conflicts, **terrorist** tactics and conventional, chemical or nuclear arms buildups that threaten the p
 ial deterrent forces to other parts of the world, toward local low-intensity conflicts and **terrorist** activities, toward hostile acts by undemocratic and unpredictable governments
 Israel forced Egypt, Syria and Jordan to proscribe direct cross-border **terrorist** attacks.
 In addition new East European governments are inhospitable to international **terrorists**, denying them indispensable training bases and logistical backing.
 which were designed for tactical use against helicopters but can also easily be used by **terrorists** against passenger-carrying commercial aircraft. 2 As the possibility, let alone lik
 Only humans, @ @ @ @ @ @ @ @ @ @ as the composition and action plans of **terrorist** groups or of narcotics traffickers.
 ickly transferring large sums of money in ways that draw @ @ @ @ @ @ @ @ @ @ by **terrorists**, narcotics traffickers and others who need to move or launder money discreetly.
 ily welcome special operations that rescued hostages and punished hostage-takers or **terrorists** for their actions, particularly if their depredations were then curtailed -- as Muar
 n being informed of the plans that the Minister of Defence had made to eliminate the " **terrorist** leaders " before be would agree to vote in favour of additional appropriations to
 ications if they had cause to believe that the aid was derived @ @ @ @ @ @ @ @ @ @ **terrorists** " in the West Bank and Gaza Strip.
 :he plan drawn @ @ @ @ @ @ @ @ @ @ his assassination by Menachem Begin's/Irgun **terrorists**), calling for a division of the old state of Palestine into Arab and Jewish sectors.
 The Terrorism Act of 1967 allows indefinite detention of those suspected of **terrorist** crimes, which would include belonging to any banned organization.
 own propaganda as right-wing support erodes for a government that negotiates with " **terrorists** " and " communists. "
 ierda Revolucionario) and was an active proponent of the human fights of suspected **terrorists**.
 The value of the Mothers is their refusal to accept the terms of the **terrorist** state, to articulate the existence of a responsible party which has disappeared t

Corpus

COCA: Academic

URN: [to be a
Metadata
Licence: ACA-

Text attribut

genre: ACAD
 subgenre: 20
 year: 1990
 source: Africa
 title: Aparthe
 Africa.
 publication in
 Vol. 37 Issu
 text id: 195
 file name: wlp
 paragraph_ty
 part of the se



Show Trend Diagram


terrorist Aa
or
+

Search within sentence



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

KWIC Statistics Map Name classification

Values: 1

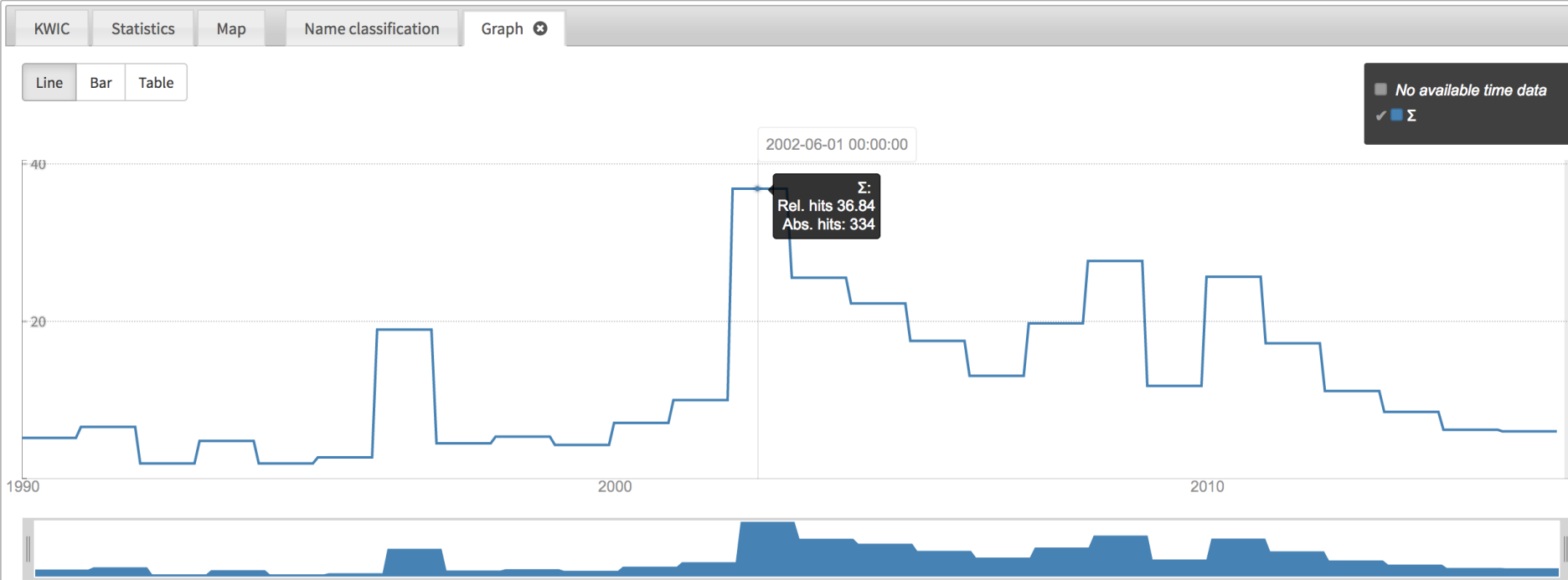
 Show Trend Diagram



<input type="checkbox"/>	word	Total	COCA: Acade...	COCA: Fictio...
<input checked="" type="checkbox"/>	Σ	 12,4 (3 029)	19,7 (2 353)	5,5 (676)
<input type="checkbox"/>	terrorist	 12,4 (3 029)	19,7 (2 353)	5,5 (676)



Trend Diagram





Korp and LAT, speech corpora

DIALECT VARIANTS OF THE WORD *METSÄ* 'FOREST'



Samples of Spoken Finnish (dialects)

ELAN 4.9.4 - SKN17a_Liperi.eaf

File Edit Annotation Tier Type Search View Options Window Help

Grid Text Subtitles Lexicon Comments Recognizers Metadata Controls

AL-sentence

Nr	Annotation
1	-- kiirettä 'malto , 'malto , "huaste'llan 'nythän se "kuul , 'samassa .
2	mittees 'sit _ossois su- , "mittees sitä 'ossois vielä , 'vielä , 'huastela ?
3	Karhusua"r _on , 'nyttiiv vie[lä] .
4	'kyllä se tuo "käuppi;as 'tietää 'onha se 'käön[yt] .
5	's _on tuossa 'yl s- , 'yl 'selän 'tuosta .
6	'tuossa on , 'tä kun _on 'semmoien , "niemimua koko , 'tä 'Ryökolahti 'kahem puolen _oj "järvet 'tuossa .

00:00:24.765 Selection: 00:00:24.765 - 00:00:27.011 2246

SKN17a_Lip...

AL-sentence	AL-sentence-norm	MP-sentence	MP-sentence-norm	N-sentence	M-sentence
Karhusua"r _on , 'nyttiiv vie[lä] .	Karhusaari on nytkin vielä	Kais siellä oli semmonenniip paikka niin kun Karhusuari ?	Kai siellä oli semmoinenkin paikka niin kuin Karhusaari		
'kyllä se tuo "käuppi;as 'tietää 'onha se 'käön[yt] .	kyllä se tuo kauppias tietää onhan se käynyt				
's _c	se or				



SKN corpus: words that begin with *metsä*



Yksinkertainen Laajennettu Edistynyt Vertailu

sana alkaa Aa
tai

Compile statistics according to original transcribed word form without diacritics

Etsi sisältä

Konkordanssi: osumia sivulla: 25 järjestä korpuksen sisällä: järjestämätön Tilastoja: laske tilastot tämän perusteella: tarkkeeton sananmuoto Näytä sanakuva Näytä kartta

Konkordanssi Tilastoja Sanakuva Kartta Nimiluokittelu

Tuloksia: 928

« < 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ... > » Siirry sivulle / 38 Näytä konteksti

SKN – SUOMEN KIELEN NÄYTEITÄ

rt siellä kuin yhden -, Kuittijärvillä kyllä paljon olen käynyt lesnoin, metsänhoitajalla luona silloin kun Venäjän puolelta hakattiin niin, piiripäällikö
kun, metsänhoitaja , tuo, metsäpartijat oli niin viisaita sitten olisi ollut niin hirmui
kun, metsänhoitaja, tuo, metsäpartijat oli niin viisaita sitten olisi ollut niin hirmuinen maksu niin ne-
aamalla sanoo lesnoi nyt se on metsäsyyniin Ouluhtiössä niin se, se se kävi sen kaupan tekemässä Venäjä
-, Ouluhtiön puita, oululainen sen - vanhan, Viikmannin, joka oli, metsänhoitajana Ouluhtiössä niin se, se se kävi sen kaupan tekemässä Venäjä
e oli, Jumalissärkistä hakattiin siinä oli kuusi hevosta porukassa ja metsään tehtiin kämppä, Pöhlönpuron varteen siellä ja, ruvettiin ja siir
metsässä .
:oolla, päähän ja toiseen, saa tehdä sen puomipuulanssin, jo saa metsän puoleiseen päähän tehdä ja alkupäähän sen puomipuu, -, pu
lanssissa oli vettä että ei kengän varret tahtonut riittää ja sitten -, metsäajoa kun niin myöhään meni kevääksi aina se ajo niin, eivät saane

Korpus

SKN – Suomen kielen näytteitä

URN: urn:nbn:fi:lb-201407141
Kuvailutiedot
Lisenssi: CC BY 4.0 (CLARIN PUB)

Tekstin ominaisuudet

otsikko: Suomussalmen murrettä
(vihko 1)
päiväys: 1978
kirjoittaja: Alpo Räisänen
paikkakunta: Suomussalmi



Statistics: transcribed forms of *metsä*

Konkordanssi

Tilastoja

Sanakuva

Kartta

Nimiluokittelu

Arvoja: 372

 Näytä trendidiagrammi

<input type="checkbox"/>	tarkkeeton sananmuoto		Yhteensä	SKN – Suomen kielen näytteitä
<input checked="" type="checkbox"/>	Σ		1 102,5 (928)	1 102,5 (928)
<input type="checkbox"/>	mettäs		72,5 (61)	72,5 (61)
<input type="checkbox"/>	mettää		61,8 (52)	61,8 (52)
<input type="checkbox"/>	metässä		42,8 (36)	42,8 (36)
<input type="checkbox"/>	mettä		41,6 (35)	41,6 (35)
<input type="checkbox"/>	metsää		27,3 (23)	27,3 (23)
<input type="checkbox"/>	mettästä		24,9 (21)	24,9 (21)
<input type="checkbox"/>	mettät		21,4 (18)	21,4 (18)
<input type="checkbox"/>	metästä		21,4 (18)	21,4 (18)
<input type="checkbox"/>	mettäst		20,2 (17)	20,2 (17)
<input type="checkbox"/>	mettäh		19 (16)	19 (16)
<input type="checkbox"/>	mettän		19 (16)	19 (16)
<input type="checkbox"/>	metsäs		19 (16)	19 (16)
<input type="checkbox"/>	mettässä		17,8 (15)	17,8 (15)
<input type="checkbox"/>	mehtää		16,6 (14)	16,6 (14)
<input type="checkbox"/>	mehtä		14,3 (12)	14,3 (12)
<input type="checkbox"/>	mettaan		13,1 (11)	13,1 (11)
<input type="checkbox"/>	mehtään		13,1 (11)	13,1 (11)
<input type="checkbox"/>	messää		11,9 (10)	11,9 (10)



Click on “Listen in Annex”

Konkordanssi: järjestä korpuksen sisällä: järjestämätön Tilastoja: Näytä sanakuva
 Näytä kartta

Konkordanssi Tilastoja Sanakuva Kartta Nimiluokittelu Konkordanssi ✕

Tuloksia: 12

« < 1 > » Siirry sivulle / 1 Näytä konteksti

SKN – SUOMEN KIELEN NÄYTEITÄ

täjä ja siellä olisi kaunis paikka mutta kun on raja niin likellä niin, kun nyt **metsä** hakattiin ja siitä tulee kolmisenkymmentä kilometriäkö täältä vaan -
enemmän kilometriäkö täältä vaan -, vieläkö enempi matkaa niin kun **metsä** hakattiin niin tämä kirkon torni selvästi näkyy.
käymään sitten jossakin tietäjässä ja se oli tiennyt sanoa että ne on niitä, **metsä** pitää.
ti että ne tänä päivänä pääsee sieltä lähtemään että ne on ollut niin kuin, **metsä** pitänyt.
onko heitä sitten, semmoisiakin tapauksia minä en tiedä pitääkö **metsä**, mutta niin se, kertoi.
no kertoikos ne sellaista että olisi ihmistäkin pitänyt **metsä** joskus?
nyt niin paikan päällä, tuosta vähän -, tuonne noin tuo -, kun tuo matala **metsä**, -, alkaa ja tuossa pitkin niin, se oli Venäjänkorpi.
eihän sitä nyt, kenenkään metsään niin se piti olla, semmoinen luvattu **metsä** mistä, meni ympäri, .
minähän lähdin sitten ja joka oli se -, **metsä** söi siltä ne vaattetkin että ne hetuleet meni vain jäljissä niin.
e kun se oli se metsänpeitto kun sanottiin että karja, karja, karjan peittää **metsä**.
jos oli äijä oksia niin se paloi - mutta jos ei ollut, oikein, sankka **metsä** ja petäjä seassa niin tuota, sitä sitten, -, rovio kun se oli pitkä rovio,
sanoin jotta en minä kehtaa kiusata kun, petäjää on kuitenkin ja, valtion **metsä** on vielä meillä tässä syötävänä niin.

« < 1 > » Siirry sivulle / 1

Lataa tiedostona muodossa:

[JSON](#)

Korpus

SKN – Suomen kielen näytteitä

URN: <urn:nbn:fi:lb-201407141>

Kuvailutiedot

Lisenssi: [CC BY 4.0 \(CLARIN PUB\)](https://creativecommons.org/licenses/by/4.0/)

Tekstin ominaisuudet

otsikko: Kalajoen murreta (vihko 19)

päiväys: 1984

kirjoittaja: Raimo Jussila

paikkakunta: Kalajoki

murrealue: Pohjalaismurteet

murreryhmä: Keski-Pohjanmaa

tiedoston nimi: SKN19b_Kalajoki

puhujat: ES

sukupuoli: M

rooli: haastateltava

Sanan ominaisuudet

alkuperäinen sana: 'mehtä

tarkkeeton sananmuoto: mehtä

kommentti: [tyhjä]

[Lataa FAF-tiedosto](#)

[Lataa teksti](#)

[Kuuntele Annexissa](#)



Listen to the utterance

- Text
- Grid
- Subtitle
- Waveform
- Timeline**
- Combined

Video display min

No video, only audio

Information min

General Session Technical

Resource: SKN19b_Kalajoki.eaf
Media file: SKN19b_Kalajoki.m4a
Elapsed time: 00:33:43:765

Selected chunk:
Begin time: 00:33:43:765
End time: 00:33:51:482
Text: -

Mini Data Frame min

Tier: none

Font size: 14

Play selection

Clear selection

Create bookmark

|< >|

<< >>

< >

+ -

Play screen by screen
 Play continually

Tier text font:
Select a font

Timeline

00:33:40:00 00:33:42:00 00:33:44:00 00:33:46:00 00:33:48:00 00:33:50:00 00:33:52:00

TT-sentence						
ES-sentence	en ny 'hänej 'jäläkiin että "hää 'c	[nc]	'määhäl "lähin 'sitte 'a joka 'oli se pit-	'mehtä "söi 'siltä ne 'vaattekki että ne "hetuleet 'meni väi "jälisä r		
TT-sentence-no						
ES-sentence-no	den nyt hänen jälkiin että hän op	[nc]	minähän lähdin sitten ja joka oli se - metsä söi siltä ne vaattetkin että ne hetuleet meni vain jäljissä niin			



Download service: www.kielipankki.fi/download



LATAUKSET
DOWNLOADS

HOME UP

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

Name	Size	Description
AMPH/	-	amph-Corpus
CEAL/	-	CEAL corpus
Digilib-pub/	-	Kansalliskirjaston lehtikokoelma 1771–1874
DSPCON/	-	Aalto University DSP Course Conversation Corpus
ELFA/	-	ELFA corpus
FBC/	-	Finnish Broadcast Corpus
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
italian-letters/	-	
KKS/	-	Karjalan kielen sanakirja (XML)
lehdet90ff/	-	Finnish Magazines and Newspapers from the 1990s and 2000s
SFNET/	-	SFNET Corpus
SNC1/	-	Kansalliskirjaston lehtikokoelman ruotsinkieliset n-grammit 1770-1940
Suomi24/	-	The Suomi 24 Corpus
UHLCS/	-	U Helsinki Language Corpus Server
Ylilauta/	-	Ylilauta Corpus



Edit / search annotations in ELAN

ELAN 4.9.1-b

ELAN - SKN19b_Kalajoki_full.eaf

File Edit Annotation Tier Type Search View Options Window Help

Text Grid Subtitles Lexicon Metadata Recognizers Comments **Controls**

Volume: 100

SKN19b_Kalajoki.wav

Mute Solo

Rate: 100

00:00:00.000 Selection: 00:00:00.000 - 00:00:00.000 0

⏮ ⏪ ⏩ ⏭ ⏮ ⏪ ⏩ ⏭ ⏮ ⏪ ⏩ ⏭ ⏮ ⏪ ⏩ ⏭

⏮ ⏪ ⏩ ⏭ ⏮ ⏪ ⏩ ⏭

Selection Mode Loop Mode

SKN19b_Kal... 00:33:42.000 00:33:43.000 00:33:44.000 00:33:45.000 00:33:46.000 00:33:47.000 00:33:48.000 00:33:49.000 00:33:50.000 00:33:51.000 00:33:52.000 00

TT-sentence [24]

ES-sentence [367] |noh| 'määhäl 'lähin 'sitte 'a joka 'oli se pit-, 'mehtä 'söi 'siltä ne 'vaattekki että ne 'hetuleet 'meni vä'j 'jälišä ni .

TT-sentence-norm [24]

TT-word-original [200]

TT-word-norm [167]

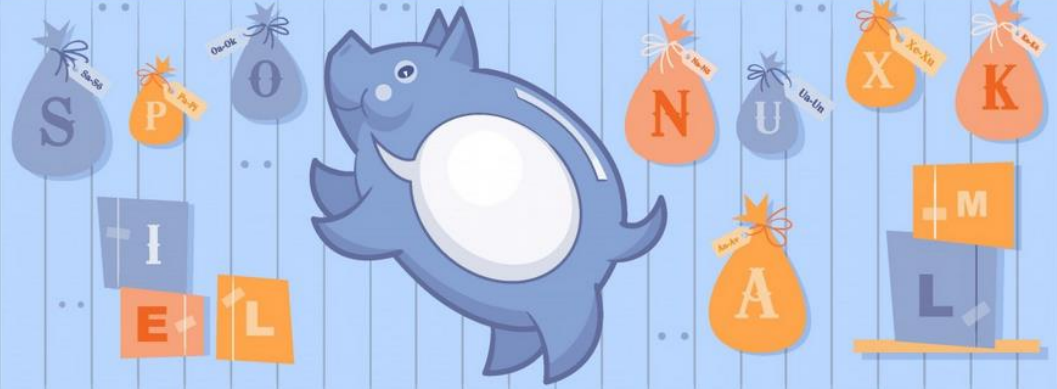
ES-sentence-norm [367] |noh| minähän lähdin sitten ja joka oli se - metsä söi siltä ne vaattetkin että ne hetuleet meni vain jäljissä niin

ES-word-original [6807] |no| 'määhäl |lähin| 'sitte |'a| joka |'oli| s |pit-|, 'mehtä |söi| 'siltä |n| 'vaattekki |että| n |hetuleet| 'meni| vä'|j| jälissä |ni| .

|no| minähän lähdin |sitten| |ia|joka| oli |s| - |metsä| söi |siltä| |n| |vaattetkin| |että| |n| |hetuleet| |meni| |vain| |jäliissä| |ni|

KIELIPANKKI

The Language Bank of Finland




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 (Download).






License

Some corpora are directly available **PUB**, some may be accessed by signing in **ACA** or by applying for individual access rights **RES**. Protected corpora can be accessed following the link  in the *Apply* column. [Access instructions](#).

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		”
aku-egg	Speech and EGG (Electroglottography) Simultaneous Recordings			LAT		”
amph	amph-Corpus			Lataus		”
ArkiSyn-korp	ArkiSyn Database of Finnish Conversational Discourse, Helsinki Korp Version			Korp		”
BeserCorp	The Corpus of Beserman Udmurt			Korp		”
	Classics of English and American Literature					”



Kielipankki corpora

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

Downloadable Version

Suomi24-2001-2014-korp	The Suomi 24 2001-2014 (Sample) Corpus, Helsinki Korp Version		Korp	”
Suomi24-2001-2015	The Suomi 24 2001-2015 (Sample) Corpus		Download	”
Suomi24-2016H2	The Suomi 24 Corpus (2016H2)		Download	”
Suomi24-korp-2016H2	The Suomi 24 Sentences Corpus (2016H2)		Korp	”



Reference 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]





Suomi24-2016H2	The Suomi 24 Corpus (2016H2)		Download	”
Suomi24-korp-2016H2	The Suomi 24 Sentences Corpus (2016H2)		Korp	”
susanne-uhlcs	The Susanne Corpus (UHLCS)		 Taito	”

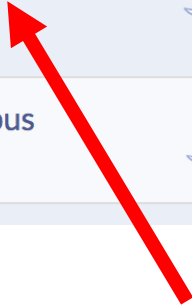




Kielipankki list of corpora

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

01-2015	(sample) Corpus				
Suomi24-2016H2	The Suomi 24 Corpus (2016H2)			Download	”
Suomi24-korp-2016H2	The Suomi 24 Sentences Corpus (2016H2)			Korp	”
susanne-uhlcs	The Susanne Corpus (UHLCS)			Taito	”



Metadata



The Suomi 24 Sentences Corpus (2016H2)

 321  16

► View resource name in all available languages

Suomi24-korp-2016H2

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

ID: <http://urn.fi/urn:nbn:fi:lb-2017021505>

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

Distribution

Availability

Available - Unrestricted Use

Licence

CC - BY - NC

Restrictions: Attribution

Attribution Details: See Documentation section.

Distribution Access/Medium: Accessible Through Interface

text

Monolingual text corpus

Languages

Finnish

Linguality

Linguality type: Monolingual

Size

2,663,114,497 Tokens

Modalities

Written Language

Metadata

Created: 09/17/2015

Last Updated: 09/17/2015

Metadata Language: English, Finnish (en, fi)

Metadata Creator

Imre Bartis 

Version






Version: 2016H2

Relation



Kielipankki list of corpora

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

SSDC-2016	Skolt Saami Documentation Corpus (2016)			LAT	”
Stu- dentsvenska	Studentsvenska 79-80 -korpus			Korp	”
Suomi24-20	The Suomi 24 2001-2014				

Licence (restricted use)



[LANGUAGE BANK](#) [ACCESS](#) [CORPORA](#) [TOOLS](#) [ORGANIZATION](#) [SUPPORT](#)

[SUOMEKSI](#) [PÅ SVENSKA](#)

CLARIN RES end-user license +PRIV 1.0

Resource: Skolt Saami Documentation Corpus (2016) (URN: urn:nbn:fi:lb-2014073037)

Copyright holder: University of Helsinki and Institute for the Languages of Finland

The Copyright holder grants the End-User a free, non-exclusive and perpetual (for the duration of the copyright) right to use and make copies of the Resource for personal use 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 user needs to be authenticated or identified.
- PLAN: The right holder requires a research plan for granting access.

General Use conditions

- BY: Attribution, i.e. acknowledgement of authorship, is required.
- PRIV: There are personal data in the resource.

Distribution conditions

- NORED: The user is not permitted to redistribute the resource.

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.



Finding corpora: Metadata catalogues





Search metadata via VLO

vlo.clarin.eu

The screenshot shows the VLO search interface. At the top, the header reads "Virtual Language Observatory" with the tagline "Explore the world of language resources and technology from different perspectives". Logos for CLARIN, IATEFL, IIR, DOBES, and others are visible. The main search area includes a search bar and a "SEARCH" button. Below the search bar, it displays "ALL RECORDS" with "889631 results" and a pagination control showing "Showing 1 to 10". Three search results are visible, each with an "Expand" button. The first result is "Deutsche Umgangssprachen: Pfeffer-Korpus" with a description of its origin at Stanford University. The second result is "CRM/" with a description of CRM resources. The third result is "Soundbites" with a description of sound fragments from the Netherlands. On the right side, there is a "NARROW DOWN" section with a list of filters: LANGUAGE, COLLECTION, RESOURCE TYPE, COUNTRY, MODALITY, GENRE, SUBJECT, FORMAT, ORGANISATION, AVAILABILITY, NATIONAL PROJECT, KEYWORD, and DATA PROVIDER. The filters LANGUAGE, COUNTRY, and ORGANISATION are highlighted with red boxes. At the bottom right of the narrow down section, there are "Expand all +" and "Collapse all -" options.



Search META-SHARE

metashare.csc.fi

The screenshot shows the META-SHARE website interface. At the top, there is an orange navigation bar with the text "META-SHARE" in the center and "Logout" on the right. Below the navigation bar, there are several menu items: "Use Resources", "Manage Resources", "Administration", "Community", "Documentation", "Statistics", and "Your Profile, Inre". The main content area features a large "META-SHARE" logo with a stylized orange and yellow graphic between the words. Below the logo, it states "311 language resources at your disposal". There is a search bar with the placeholder text "Type in your keywords, please..." and a yellow "Search" button. At the bottom left, there is a colorful graphic of speech bubbles. To the right of the graphic, there is a section titled "What is it? - About the project" with a brief description: "META-SHARE, the open language resource exchange facility, is devoted to the sustainable sharing and dissemination of language resources (LRs) and aims at increasing access to such resources in a global scale."



Language Bank support

www.kielipankki.fi/support

Helpdesk regarding the corpora:
Technical support:

fin-clarin@helsinki.fi

kielipankki@csc.fi

https://www.kielipankki.fi/tuki/

KIELIPANKKI
The Language Bank of Finland

KIELIPANKKI KÄYTTÄJÄKSI AINEISTOT TYÖKALUT FOORUMI ORGANISAATIO TUKI IN ENGLISH

Ohjeita ja oppaita

- Käyttöoikeudet ja niiden hakeminen
- Tutkijan käyttöliittymä
- Työkalujen käyttöohjeet

Uutisia



Courses

- FIN-CLARIN organizes online courses that are open to all Finnish universities (restricted number of participants from abroad):
 - Corpus linguistics and statistical methods (introductory course), 5 credits
 - Puheen analyysin perusteet – Introduction to speech analysis (Praat and ELAN), 5 credits
 - Corpus Clinic – Aineistoklinikka, 5 credits
- Starting from autumn 2018, Introduction to speech analysis will be organized in English.



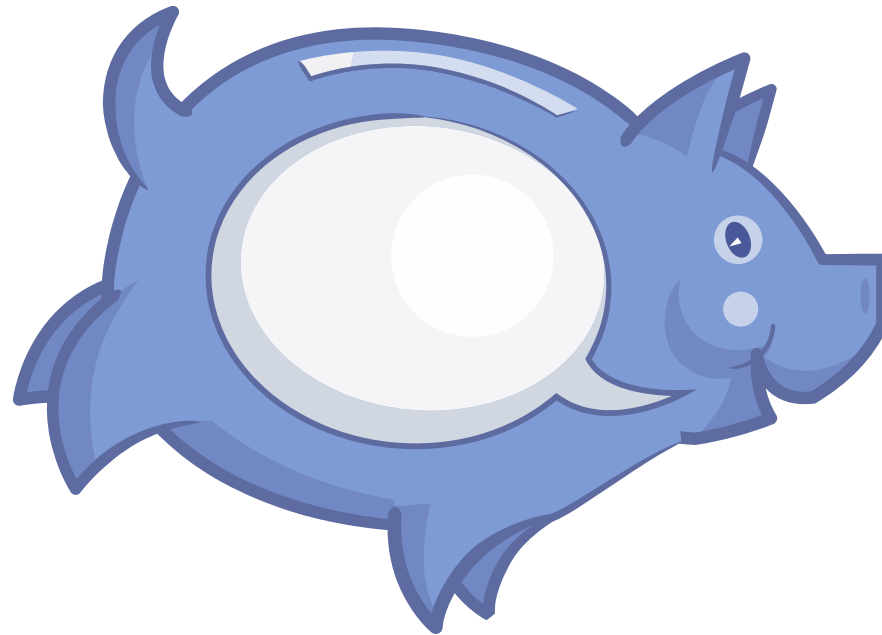
Staying tuned

- Subscribe to the newsletter (see website)
- Follow us on
 - Facebook: Kielipankki – The Language Bank of Finland
 - Twitter: @FinClarín
- Website:

www.kielipankki.fi

KIELIPANKKI

The Language Bank of Finland



Kiitos! Thank you!