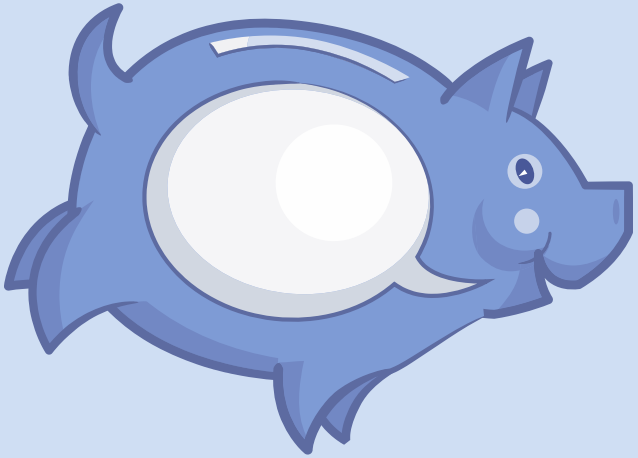


KIELIPANKKI
The Language Bank of Finland

Kielipankki – The Language Bank of Finland

Resources and services provided by FIN-CLARIN

Mietta Lennes, University of Helsinki

FIN-CLARIN

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.

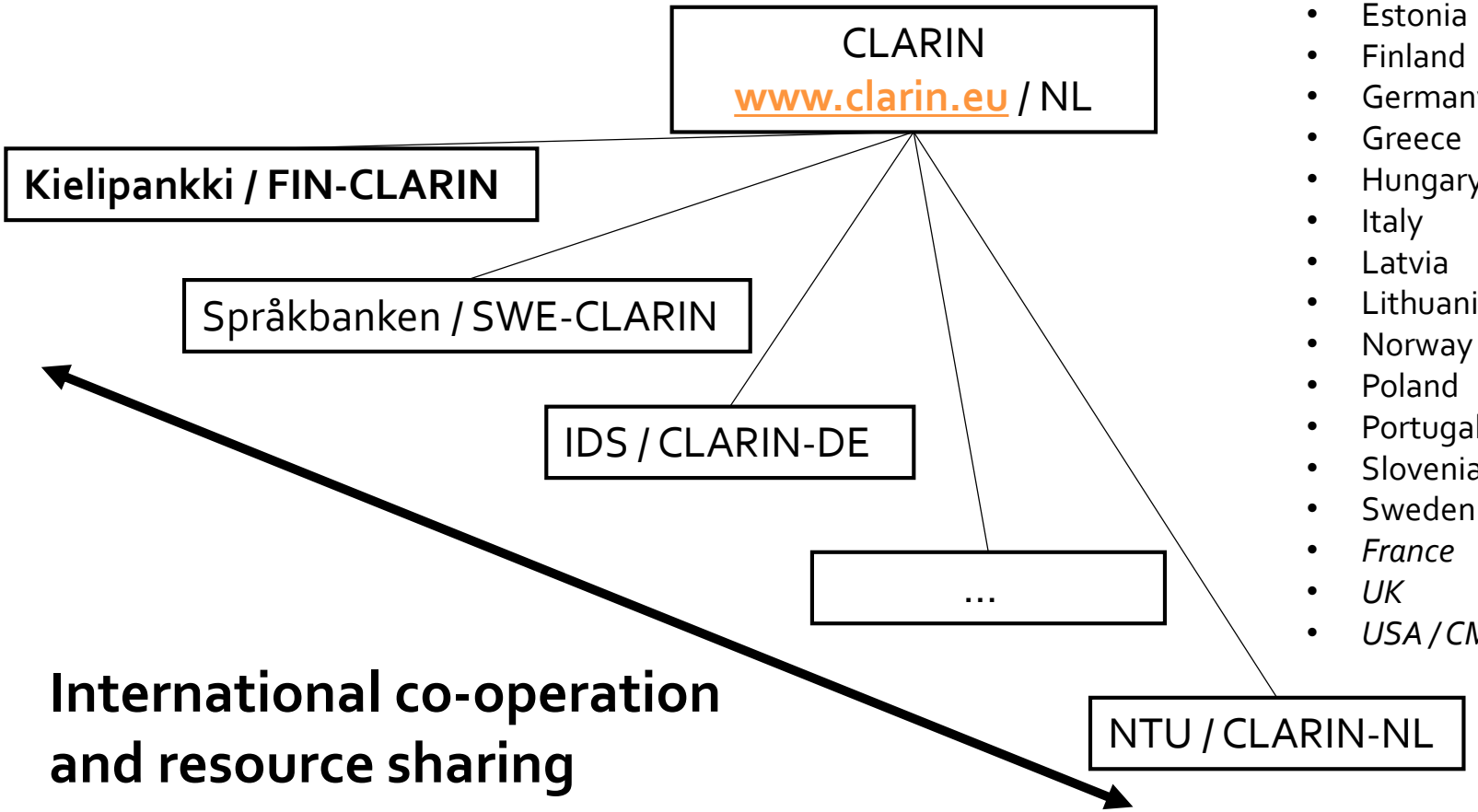
Researcher of the Month: Eero Voutilainen

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*

International co-operation and resource sharing

KIELIPANKKI
The Language Bank of Finland

CLARIN
Common Language Resources and Technology Infrastructure

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**

FIN-CLARIN

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.

2014 DSA 2017

CLARIN CENTRE B

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- (Suomi24, 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 an b publ.)	12 Gw	20 Gw
Social media and similar sources 2000- (, Ylilauta, …)	4 Gw	10 Gw
Literature and manuscripts (Gutenber rchives)	60 Mw	70 Mw
Speech		
News broadcasts (YLE)		10000 h
Video sessions from the Finnish P	500 h	1000 h
Dialect and everyday speech (Ko	500 h	1000 h
Sign language resources (Aalto lintto)	20 h	500 h
Multilingual and Other Resou		
Multilingual Resources (Euro arl, 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 juntan	terroristit	aloittivat yöllä 02.00 ensin tulituksen vapausarm
yöllä 02.00 ensin tulituksen vapausarmeijan asemiin ja sen jälkeen	terroristien	jalkaväki käyi hyökkäyksen

Corpus

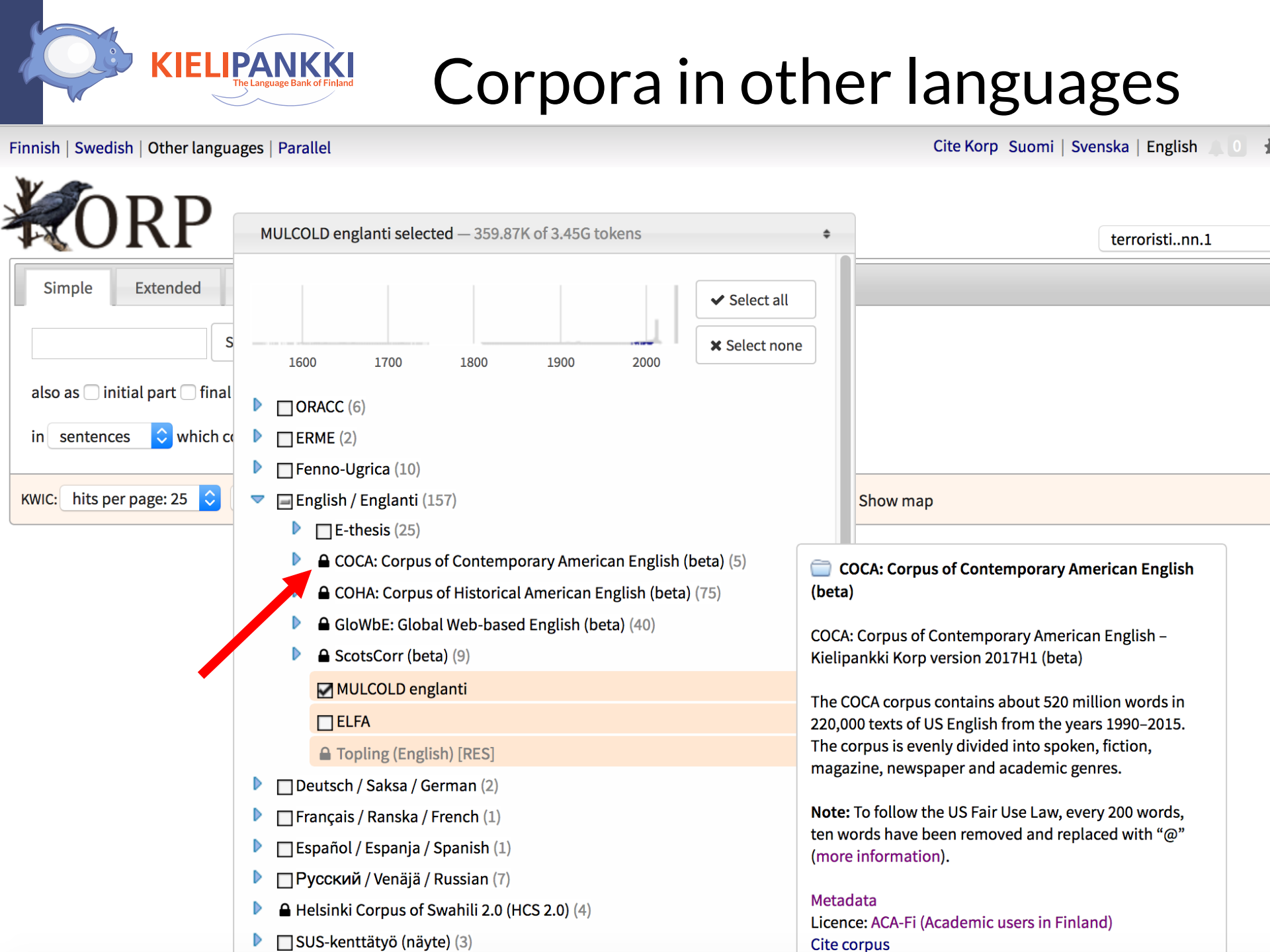
Suomi24 (1/10)

Metadata
Licence: CC BY-NC (CLARIN PUB)
Cite corpus
Link to corpus in Korp: urn:nbn:fi:lb-2015120401

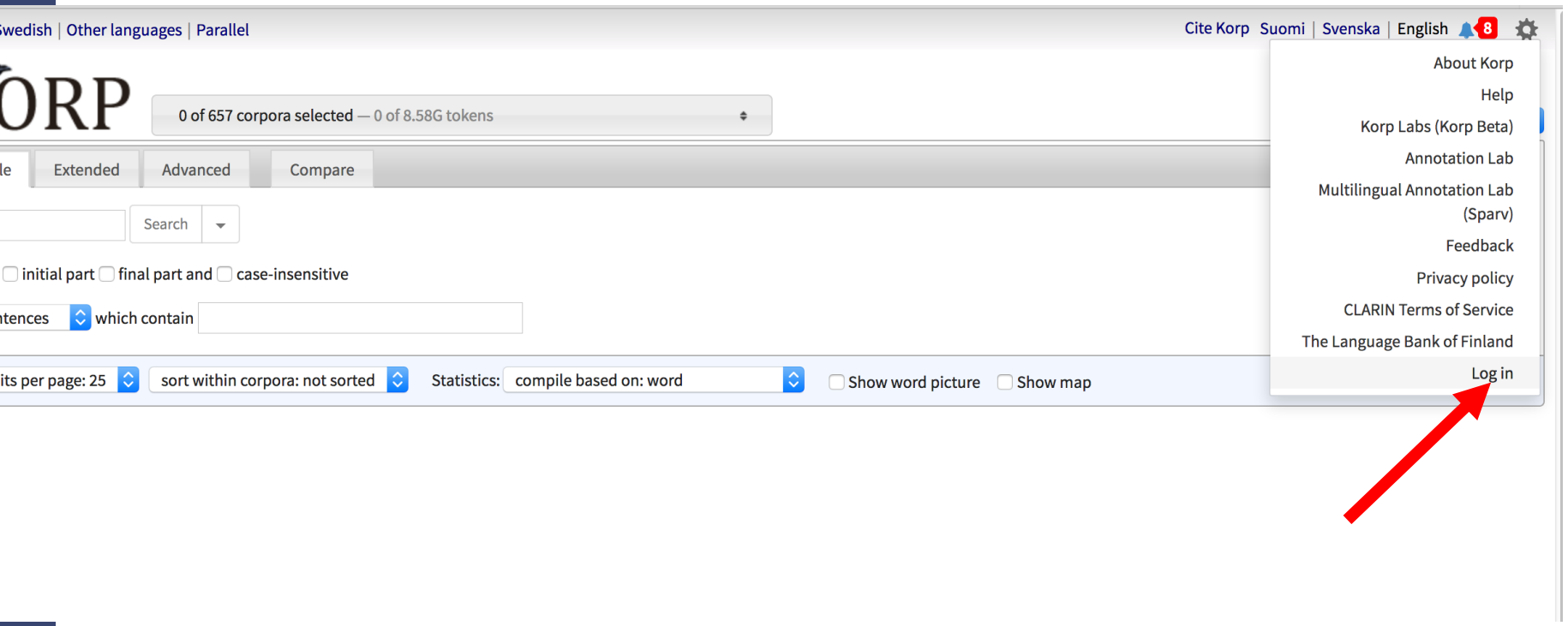
Text attributes

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

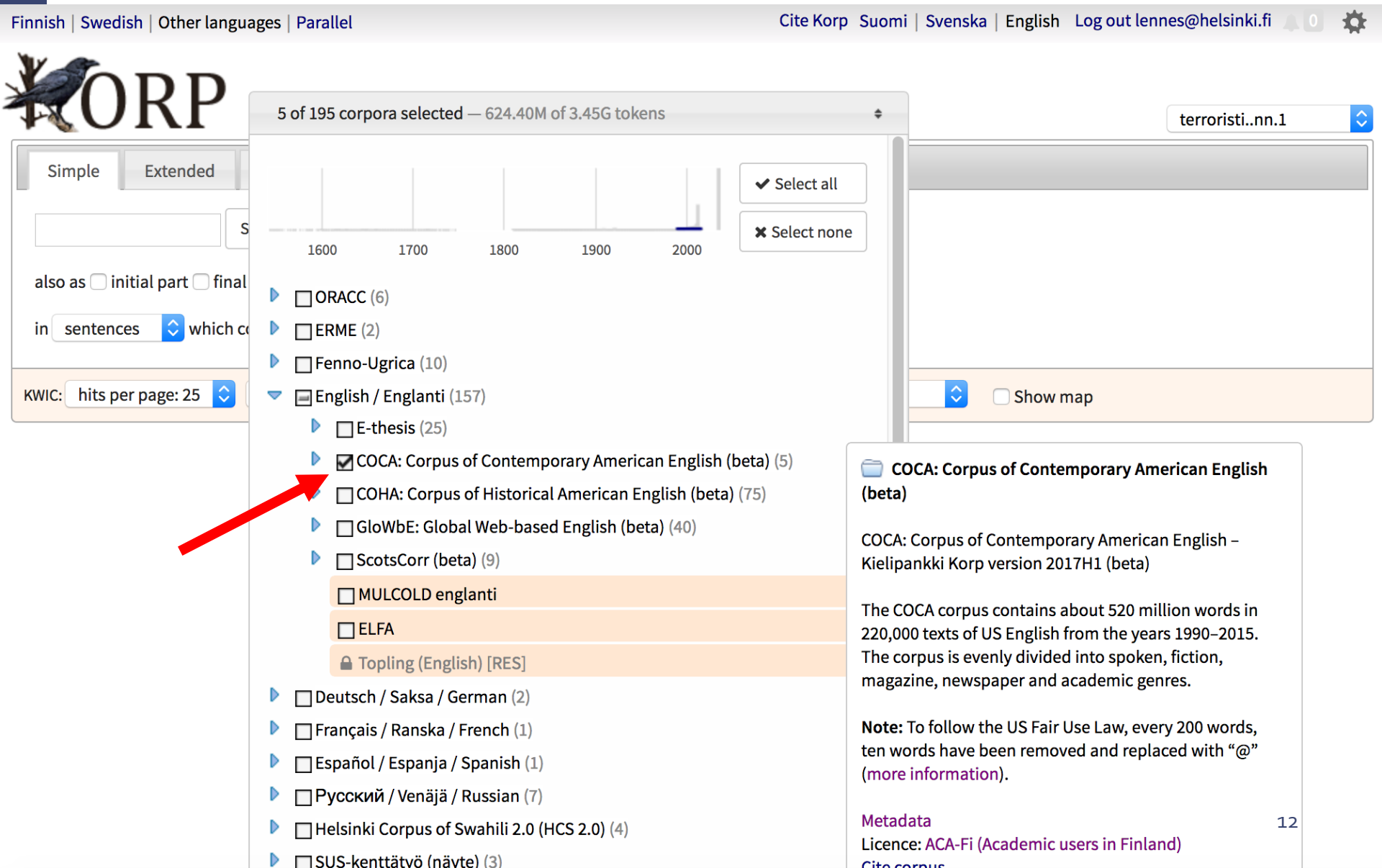
Corpora in other languages



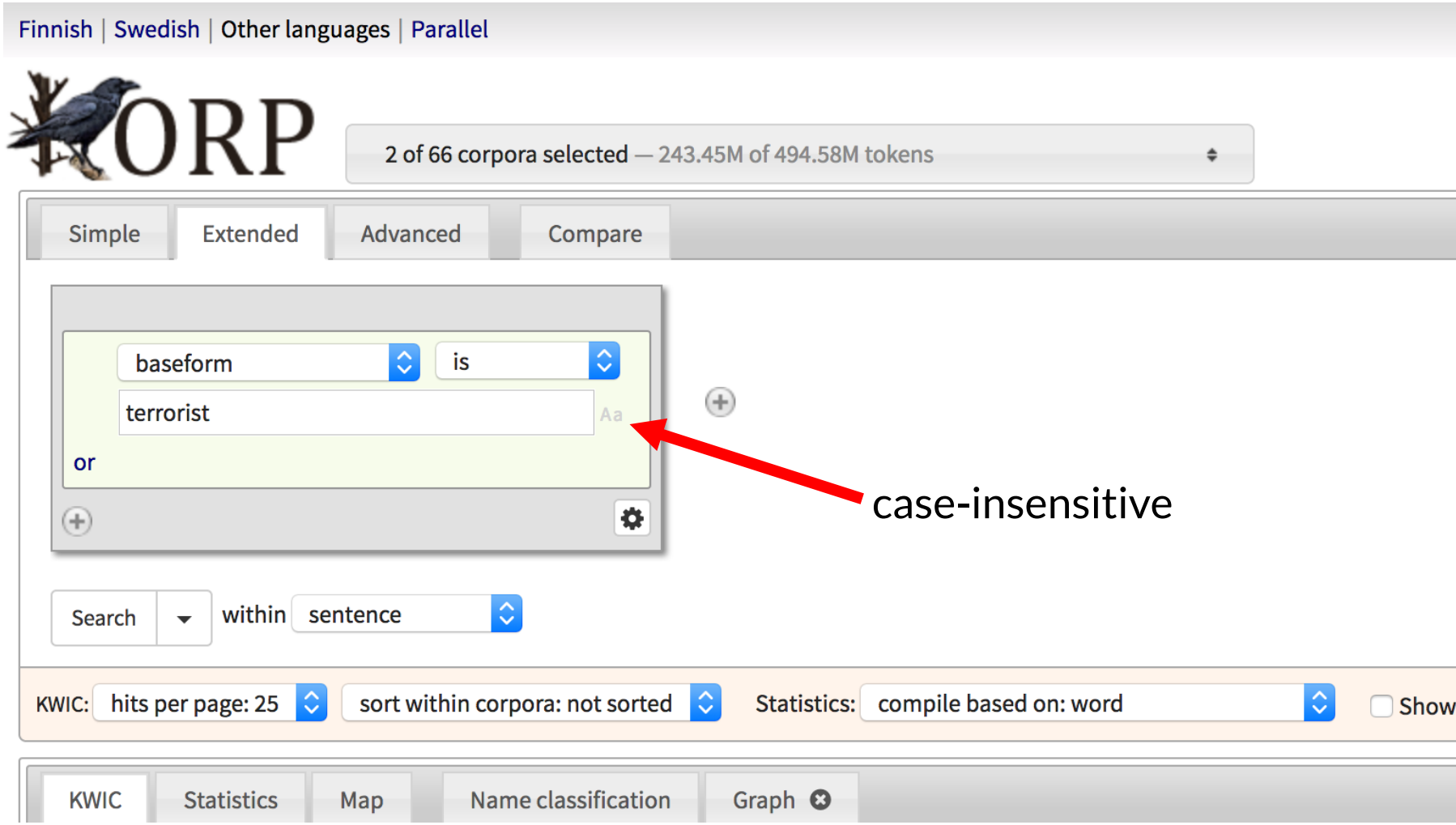
Log in to use an ACA-licensed corpus



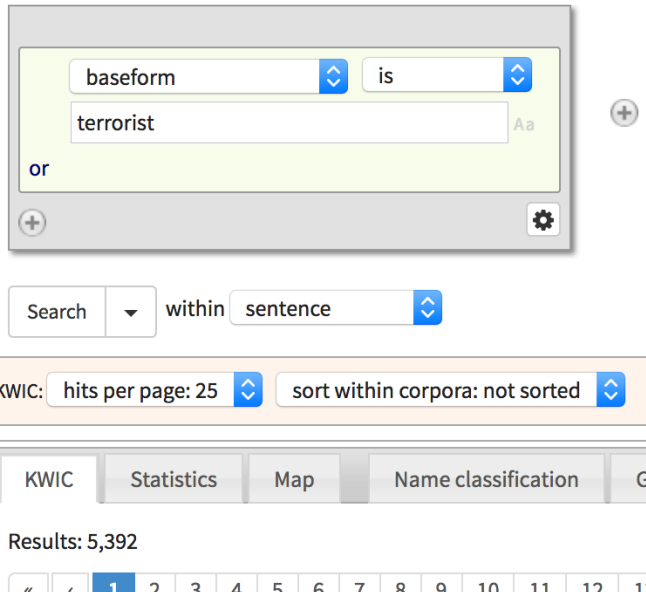
Now select the COCA corpus in the corpus menu



Search for any form of the word ‘terrorist’ (“Extended” tab)



’terrorist’


baseform
is
terrorist
Aa
or
Search
within
sentence
KWIC:
hits per page: 25
sort within corpora: not sorted
Statistics:
compile based on: word
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
ct the rights of @ @ @ @ @ @ @ @ @ kind of regional, racial and religious conflicts, terrorist tactics and conventional, chemical or nuclear arms buildups that threaten the p
al 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.
uickly transferring large sums of money in ways that draw @ @ @ @ @ @ @ @ @ @ by terrorists, narcotics traffickers and others who need to move or launder money discreetly.
ly 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
ciations 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 th
Corpus
COCA: Acade
URN: [to be a
Metadata
Licence: ACA-
Text attribut
genre: ACAD
subgenre: 20
year: 1990
source: Africa
title: Aparthe
Africa.
publication i
Vol. 37 Issu
text id: 195
file name: wl
paragraph_ty
part of the se

Show Trend Diagram

Search within sentence

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

KWIC Statistics Map Name classification

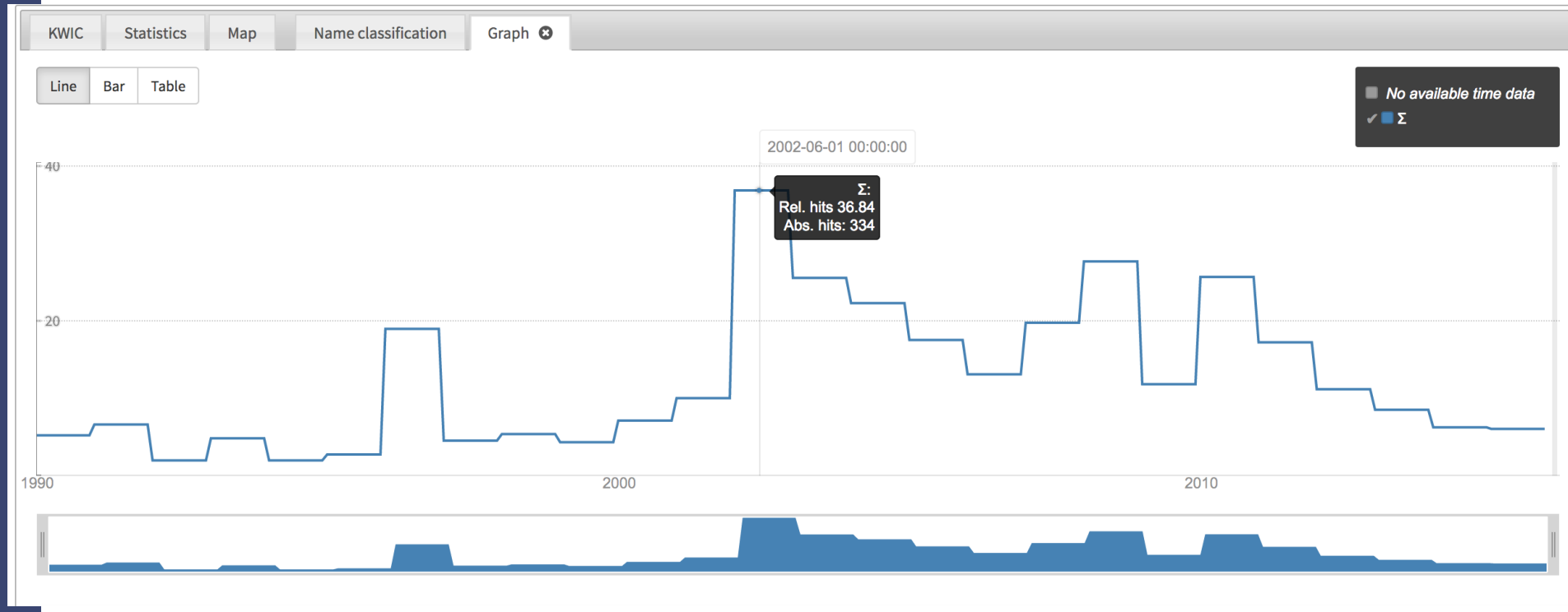
Values: 1

Show Trend Diagram

word	Total	COCA: Acade…	COCA: Fictio…
Σ	12,4 (3 029)	19,7 (2 353)	5,5 (676)
terrorist	12,4 (3 029)	19,7 (2 353)	5,5 (676)

KIELIPANKKI
The Language Bank of Finland

Trend Diagram



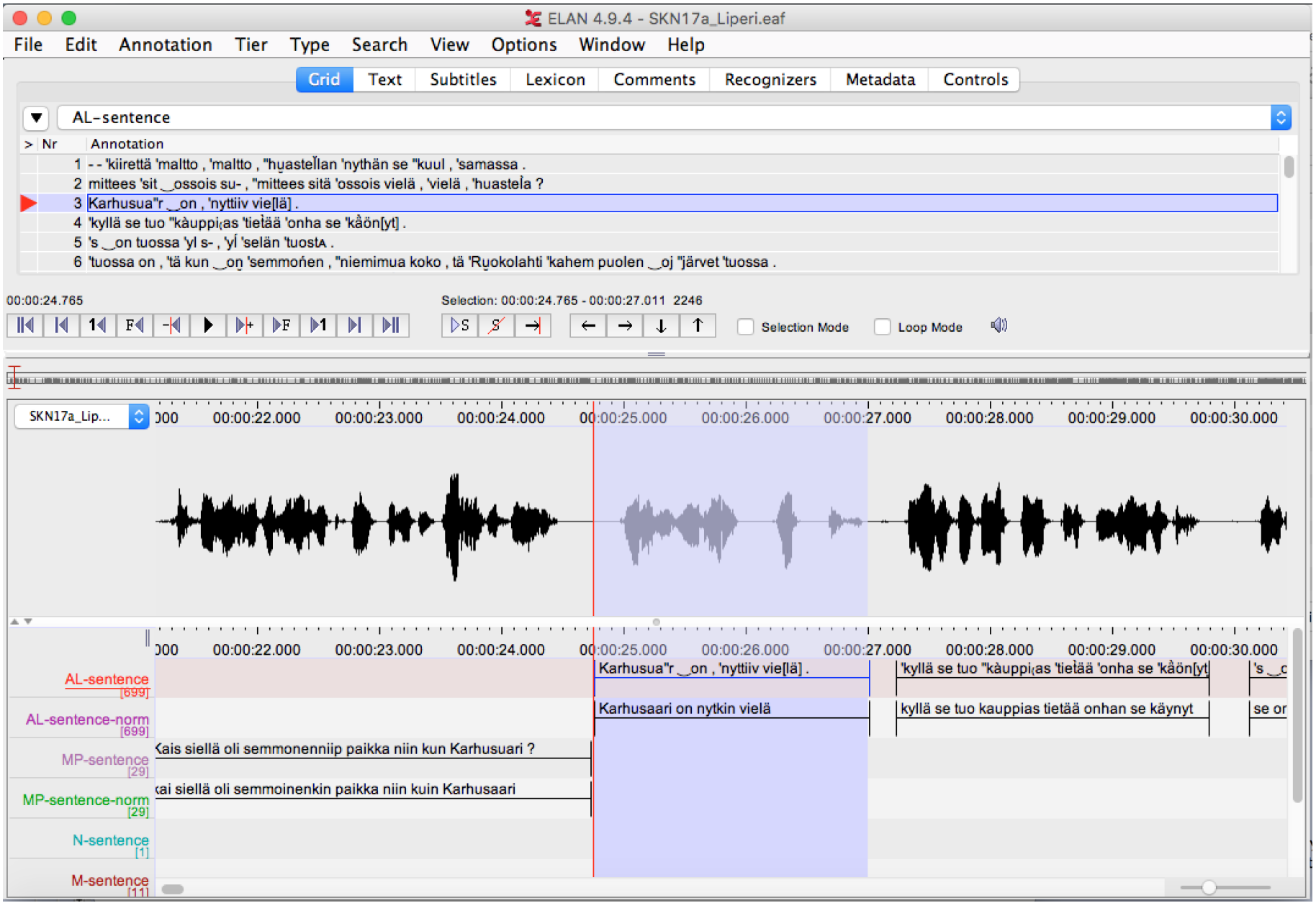
FIN-CLARIN

Korp and LAT, speech corpora

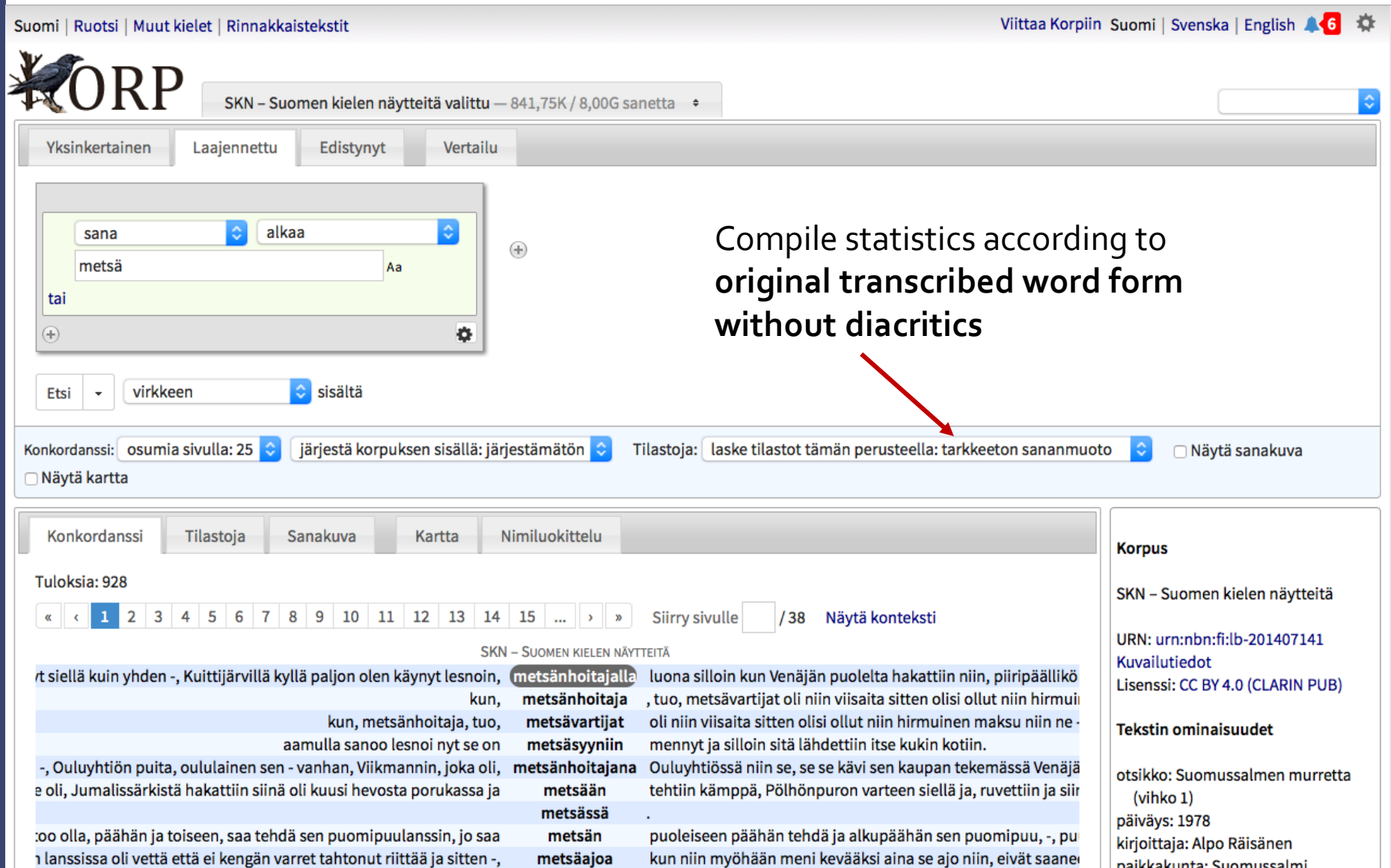
DIALECT VARIANTS OF THE WORD *METSÄ* 'FOREST'

KIELIPANKKI
The Language Bank of Finland

Samples of Spoken Finnish (dialects)



SKN corpus: words that begin with *metsä*



Statistics: transcribed forms of *metsä*

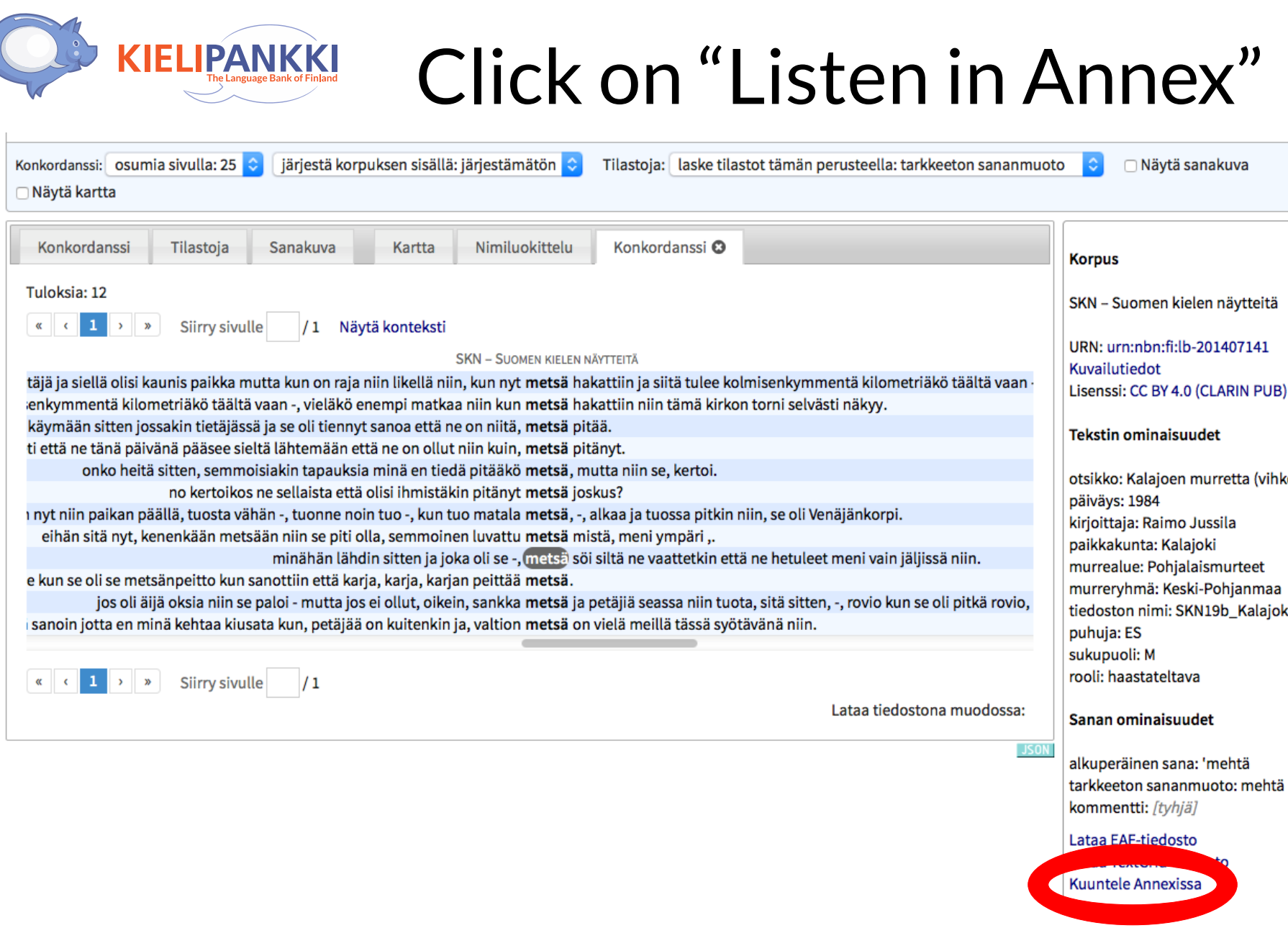
Konkordanssi | Tilastoja | Sanakuva | Kartta | Nimiluokittelu

Arvoja: 372

Näytä trendidiagrammi

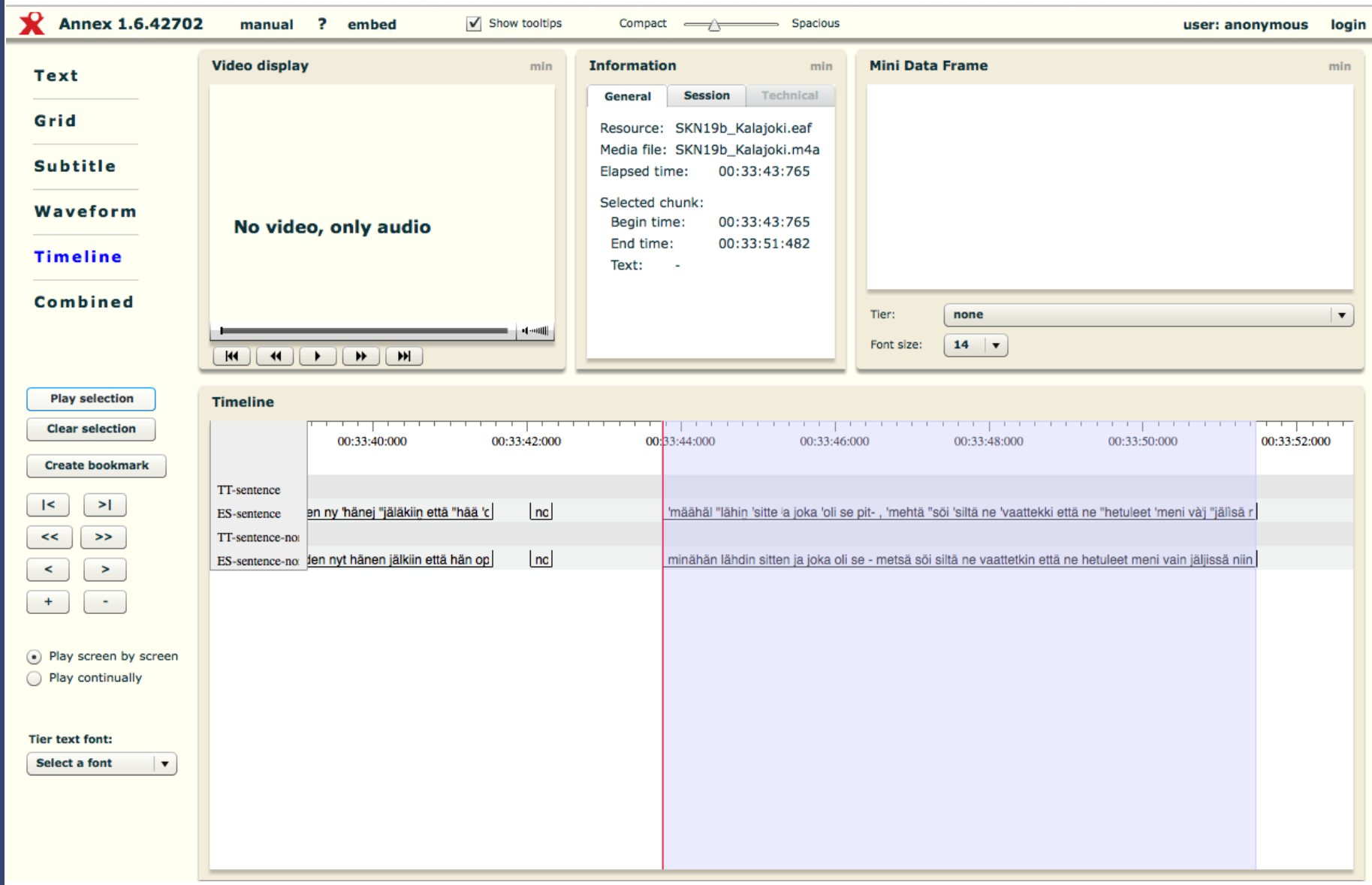
tarkkeeton sananmuoto	Yhteensä	SKN – Suomen kielen näytteitä
Σ	1 102,5 (928)	1 102,5 (928)
mettäs	72,5 (61)	72,5 (61)
mettää	61,8 (52)	61,8 (52)
metässä	42,8 (36)	42,8 (36)
mettä	41,6 (35)	41,6 (35)
metsää	27,3 (23)	27,3 (23)
mettästä	24,9 (21)	24,9 (21)
mettät	21,4 (18)	21,4 (18)
metästä	21,4 (18)	21,4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

Click on “Listen in Annex”



KIELIPANKKI
The Language Bank of Finland

Listen to the utterance



Download service: www.kielipankki.fi/download

KIELIPANKKI
The Language Bank of Finland

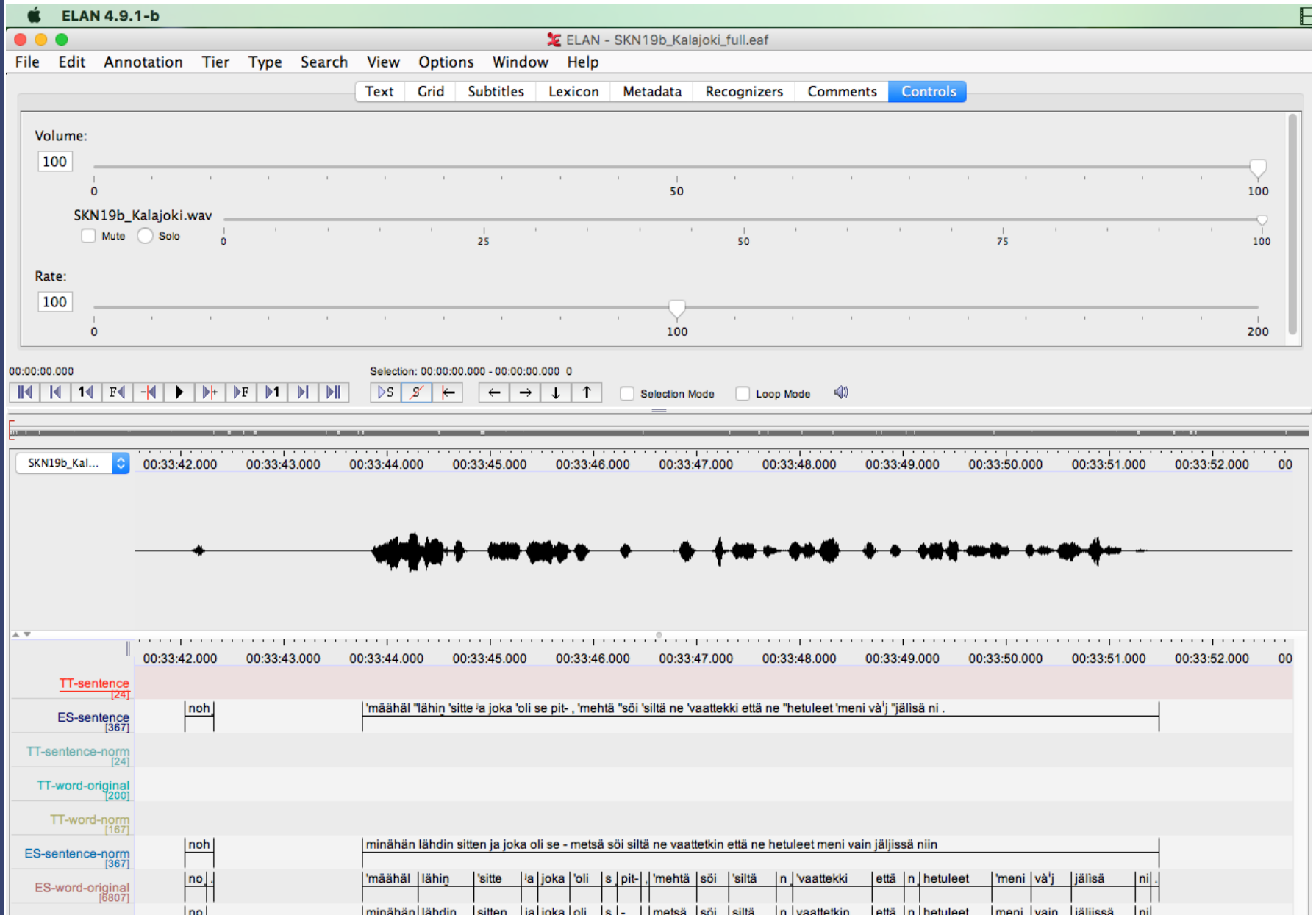
LATAUKSET
DOWNLOADS

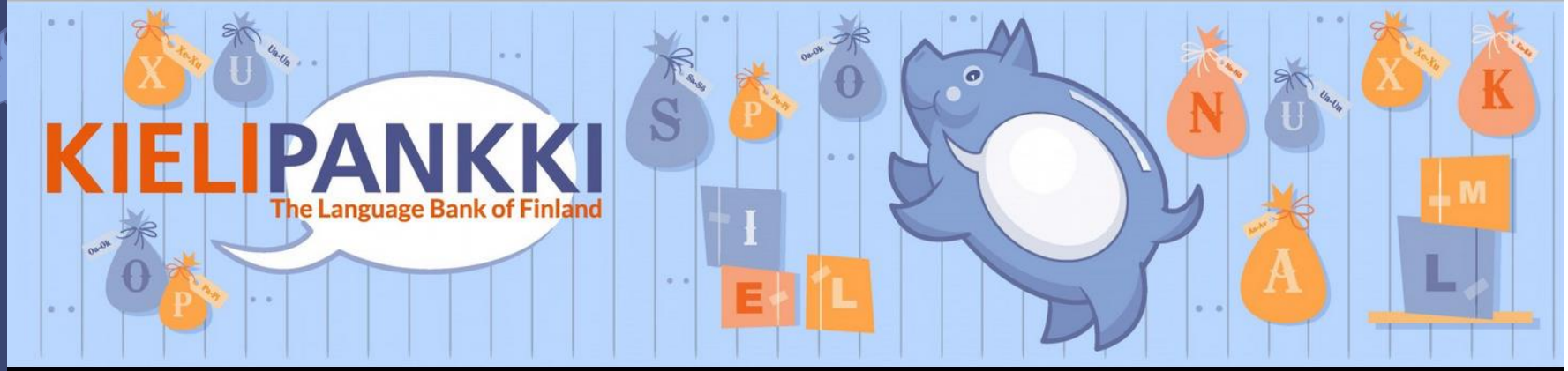
HOME UP

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

Name	Size	Description
AMPH/	-	amph-Corpus
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





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	PUB		Korp		❞
aku-egg	Speech and EGG (Electroglottography) Si-multaneous Recordings	ACA		LAT		❞
amph	amph-Corpus	ACA		Lataus		❞
ArkiSyn-korp	ArkiSyn Database of Finnish Conversational Discourse, Helsinki Korp Version	PUB		Korp		❞
BeserCorp	The Corpus of Beserman Udmurt	PUB		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	PUB	Korp	❞
Suomi24-2001-2015	The Suomi 24 2001-2015 (Sample) Corpus	ACA	Download	❞
Suomi24-2016H2	The Suomi 24 Corpus (2016H2)	ACA	Download	❞
Suomi24-korp-2016H2	The Suomi 24 Sentences Corpus (2016H2)	PUB	Korp	❞

Reference instructions

Kielipankki corpora

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

[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)	ACA		Download	❞
Suomi24-korp-2016H2	The Suomi 24 Sentences Corpus (2016H2)	PUB		Korp	❞
susanne-uhlcs	The Susanne Corpus (UHLCS)	RES	●	Taito	❞

Kielipankki list of corpora

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

01-2015	(Sample) Corpus			
Suomi24-2016H2	The Suomi 24 Corpus (2016H2)	ACA	Download	99
Suomi24-korp-2016H2	The Suomi 24 Sentences Corpus (2016H2)	PUB	Korp	99
susanne-uhlcs	The Susanne Corpus (UHLCS)	RES	Taito	99

Metadata

The Suomi 24 Sentences Corpus (2016H2)

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

text

Distribution

Availability

Available - Unrestricted Use

Licence

CC - BY - NC

Restrictions: Attribution

Attribution Details: See Documentation section.

Distribution Access/Medium: Accessible Through Interface

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)	RES		LAT	
Stu-dentsvenska	Studentsvenska 79-80 -korpus	RES		Korp	
Suomi24-20	The Suomi 24 2001-2014				

Licence (restricted use)

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.

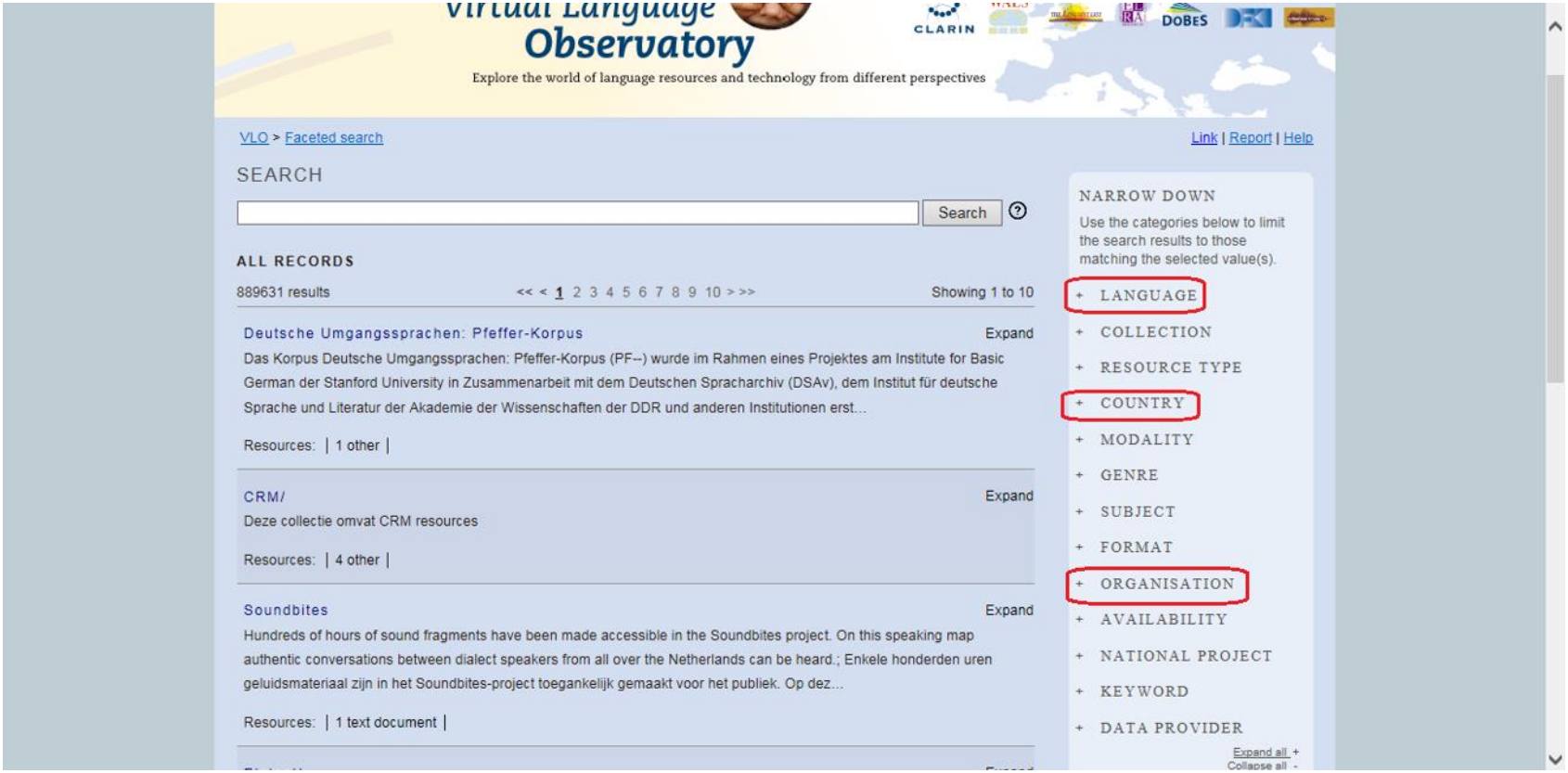
This license has been made in compliance with copyright agreements by WIPO – the World Intellectual Property Organization. The rights granted in this license shall be so interpreted that in case applicable intellectual property laws grant rights not mentioned in this license, they are also regarded as 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

KIELIPANKKI
The Language Bank of Finland

Search metadata via VLO

vlo.clarin.eu



FIN-CLARIN

KIELIPANKKI
The Language Bank of Finland

Search META-SHARE

metashare.csc.fi

FIN-CLARIN

Language Bank support

www.kielipankki.fi/support

Helpdesk regarding the corpora: fin-clarin@helsinki.fi
Technical support: kielipankki@csc.fi

KIELIPANKKI
The Language Bank of Finland

Courses

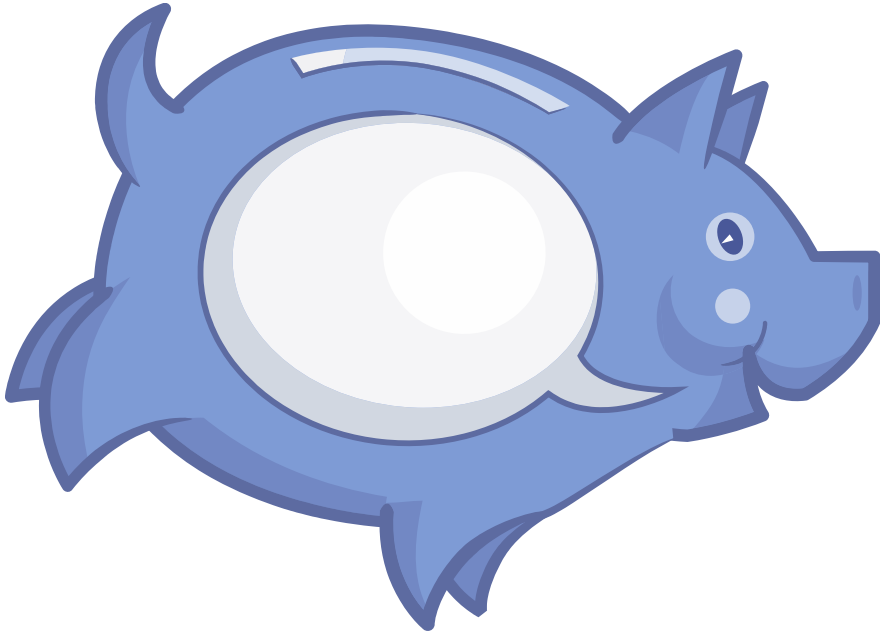
- FIN-CLARIN organizes online courses that are open to all Finnish universities (restricted number of participants from abroad):
  - Corpus linguistics and statistical methods (introductory course), 5 credits
  - Puheen analyysin perusteet – Introduction to speech analysis (Praat and ELAN), 5 credits
  - 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: @FinClarin
- Website:

www.kielipankki.fi

KIELIPANKKI
The Language Bank of Finland



Kiitos! Thank you!