

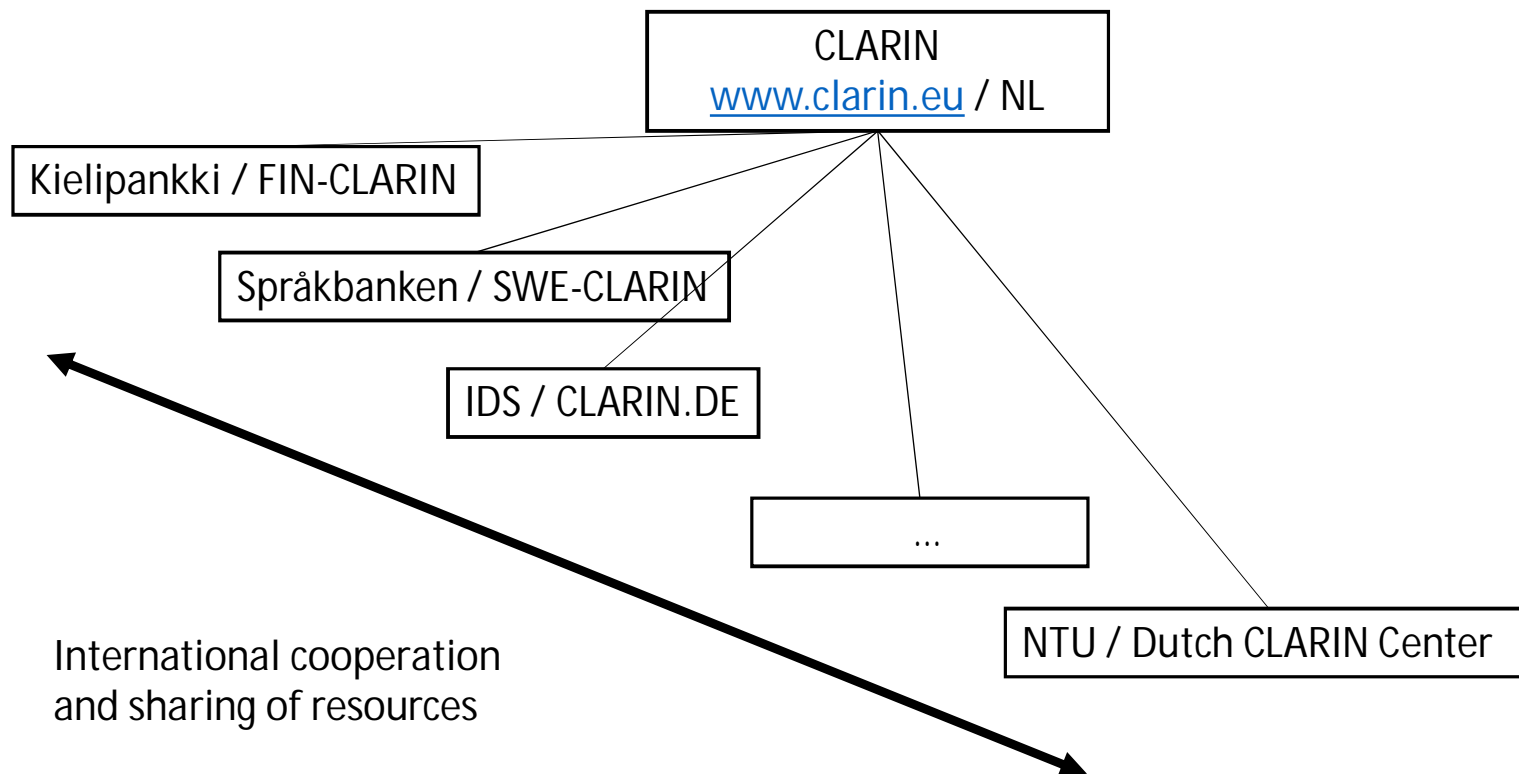
# FIN-CLARIN and CLARIN

Infrastructure for Digital Humanities

Krister Lindén, National Coordinator of FIN-CLARIN

# CLARIN ERIC

European Research Infrastructure Consortium  
founded on February 29, 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

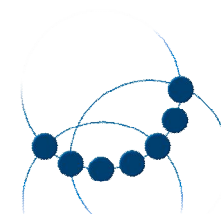
[www.kielipankki.fi](http://www.kielipankki.fi):

- University of Helsinki
- CSC – IT Center for Science
- 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

Coordinate the activity and provide access to large centrally acquired resources and tools

Provide access to resources and tools developed locally by individual researchers or research groups

**CLARIN**



Common Language Resources and Technology Infrastructure

# 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

Learn more

Search



Showing all 1617743 records

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 > >>

## EXMARaLDA Demo corpus

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

A selection of short audio and video recordings in various languages to be used for instruction or demonstration of the EXMARaLDA system.; HIAT (simplified); HIAT; free comment; suprasegmental information; accentuation/stress; English translation; Standard German translation; German translation; English translation; code-switch



## The Hamburg MapTask Corpus (HAMATAC)

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

Audio and two video recordings of map tasks with adult L2 users of German and one L1 speaker. The speakers' L1 and their L2 proficiencies vary. The maps used for the tasks are available.; orthographic transcription/simplified HIAT; Fine-grained part of speech tagging using TreeTagger and the STTS tagset.; superordinate



# FIN-CLARIN Corpora for access or download

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

Resources	2017	2022
<i>Text</i>		
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
<i>Speech</i>		
News broadcasts (YLE)		10000 h
Video sessions from the Finnish Parliament 2008-2016	500 h	1000 h
Dialect and everyday speech (Kotus, Turku)	500 h	1000 h
Sign language resources (Aalto, Kuurojen liitto)	20 h	500 h
<i>Multilingual and Other Resources</i>		
Multilingual Resources (EuroParl, laws, Bible, subtitles, ...)	3 Gw	10 Gw
Learner's resources (Oulu, Jyväskylä, Kotus, Aalto)	2 Mw	5 Mw
Open source lexicons and terminologies (Helsinki, Tromssa)	300 Kw	400 Kw

Currently,  
FIN-CLARIN has  
approx. 18 GW  
in >650 databases



LANGUAGE BANK ACCESS CORPORA TOOLS FORUM ORGANIZATION SUPPORT SUOMEKSI PÅ SVENSKA

#### Access



Apply for rights to use our language resources.

#### Corpora



Browse our corpora.

#### Tools



Try our tools.

#### Forum



Talk to other users and administrators.

#### Organization



Who are the Language Bank?

#### Support



Help and instructions.



Researcher of the Month: Markus Juutinen

#### News

- Researcher of the Month: Markus Juutinen (10.10.2017)
- Network maintenance 10.10.2017 6:30am-9:00am (27.9.2017)
- Researcher of the Month: Laura-Maija



# Text, Speech and Lexical data

☐ Suomenkielinen Gutenberg -korpus

☐ Suomalaisen kirjallisuuden klassikoita (näyte)

☐ Aleksis Kivi (SKS)

☐ SKVR

▶ ☐ Juridisia tekstejä (3)

▶ ☐ Internet-keskusteluaineistoja (12)

▶ ☐ 1990- ja 2000-luvun suomalaisia aikakaus- ja sanomalehtiä (52)

▶ ☐ Muita tekstejä (11)

▼ ☐ Puhuttua kieltä (tekstiksi litteroituna) (136)

▼ ☐ Lauseopin arkiston murrekorpus (133)

▶ ☐ Lounaismurteet (20)

▶ ☐ Lounaiset välimurteet (14)

▶ ☐ Hämmäläismurteet (25)

▶ ☐ Pohjalaismurteet (27)

▶ ☐ Savolaismurteet (32)

▶ ☐ Kaakkoismurteet (15)

☐ Sananparsikokoelma (näyte)

☐ SKN – Suomen kielen näytteitä

☐ DMA – Digitaalinen muoto-opin arkisto

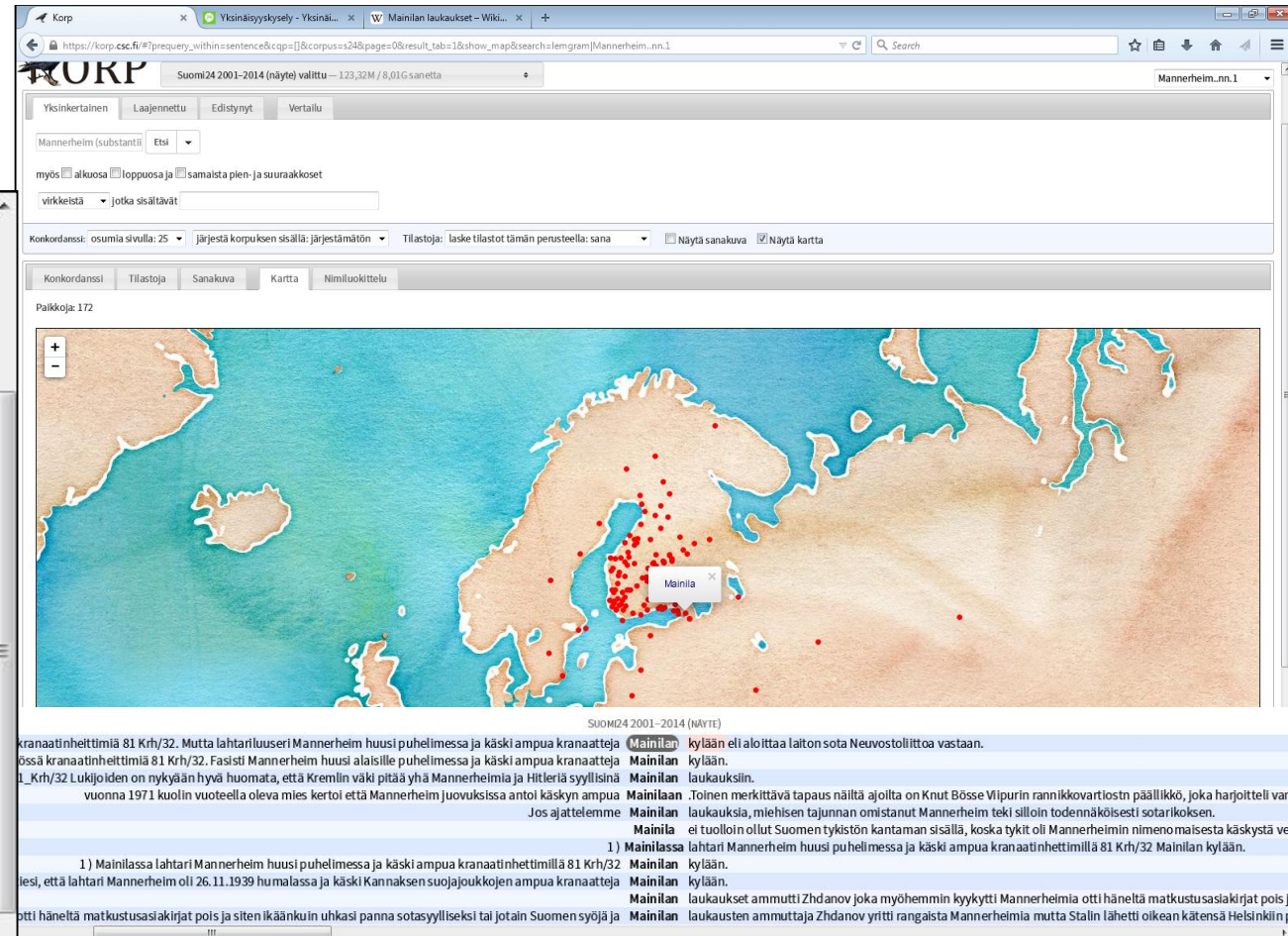
▶ ☐ Suomenoppiloiden kieltä (suomi toisena tai vieraana kielenä) (3)

▼ ☐ Vanhan kirjasuomen korpus (12)

☐ Mikael Agricolan teoksia

☐ Biblia 1642

☐ Lakeja ja asetuksia 1500–1810



## Visualizations of search results

Yksinkertainen Laajennettu Edistynyt Vertailu

perusmuoto on  
köyhäinhoito Aa  
perusmuoto on  
sosiaalihuolto Aa  
tai

# Trend Diagrams

Etsi virkkeen sisältä

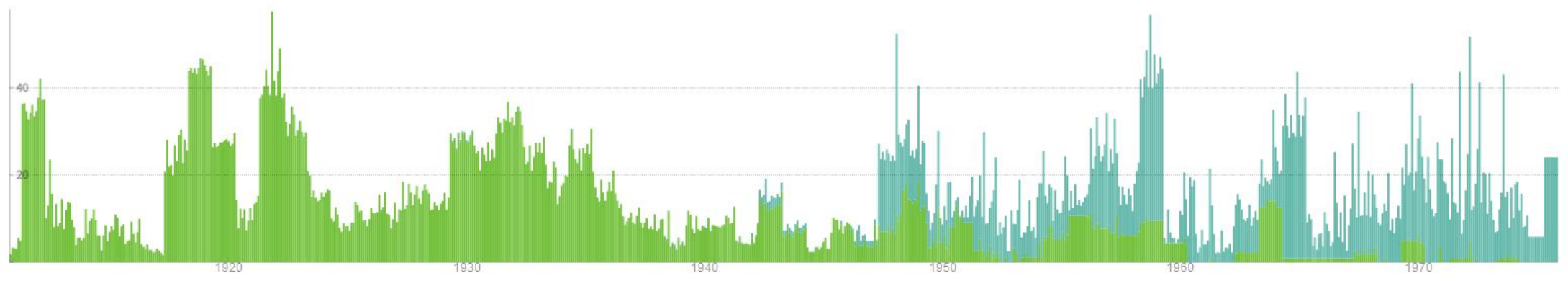
Konkordanssi: osumia sivulla: 25 järjestä korpuksen sisällä: järjestämätön Tilastoja: laske tilastot tämän perusteella: perusmuoto Näytä sanakuva Näytä kartta

Konkordanssi Tilastoja Sanakuva Kartta Nimiluokittelu Kuvaaja

Viiva Pylväs Taulukko

Aikatieto puuttuu 0.03%:sta valitusta aineistosta

Ajanjaksolta ei ole tietoja  
✓ sosiaalihuolto  
✓ köyhäinhoito






# Links to audio and video data

02 manual ? embed ☒ Show tooltips Compact  Spacious

**Video display** min



0:19 / 1:04

Full Buffer

**Information** min

General Session Technical

Resource: reitti\_a-siipeen.eaf  
Media file: reitti\_a-siipeen.mp4  
Elapsed time: 00:00:18:707

Selected chunk:  
Begin time: 00:00:18:707  
End time: 00:00:19:706  
Text: -

**Mini Data Frame**

Tier: none

Font size: 14 ▼

**Timeline**

00:00:17:500 00:00:17:750 00:00:18:000 00:00:18:250 00:00:18:500 00:00:18:750 00:00:19:000 00:00:19:250 00:00:19:500 00:00:19:750

ML-utterance	joo	
TA-utterance		eli okei mä oon tolla pualella
noise		
ML-word	joo	
TA-word		eli okei mä oo tolla pualella
ML-word-norm	joo	
TA-word-norm	haa	eli okei minä ole tuolla puolella
TA-gesture		stroke hold

## Links to audio and video data

02 manual ? embed ☒ Show tooltips Compact  Spacious

**Video display** min



0:19 / 1:04

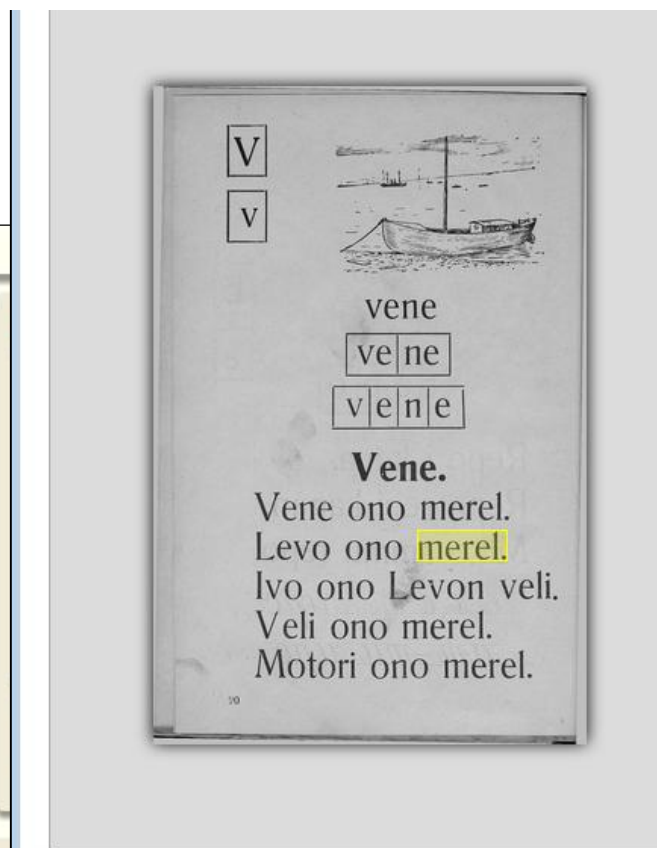
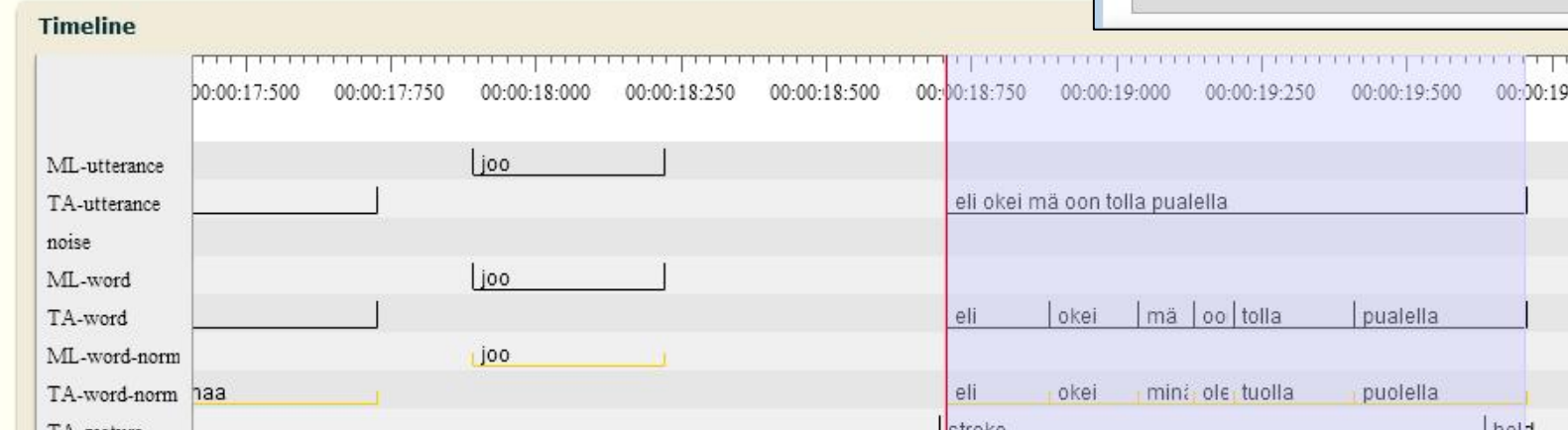
Full Buffer

**Information** min

General Session Technical

Resource: reitti\_a-siipeen.eaf  
Media file: reitti\_a-siipeen.mp4  
Elapsed time: 00:00:18:707

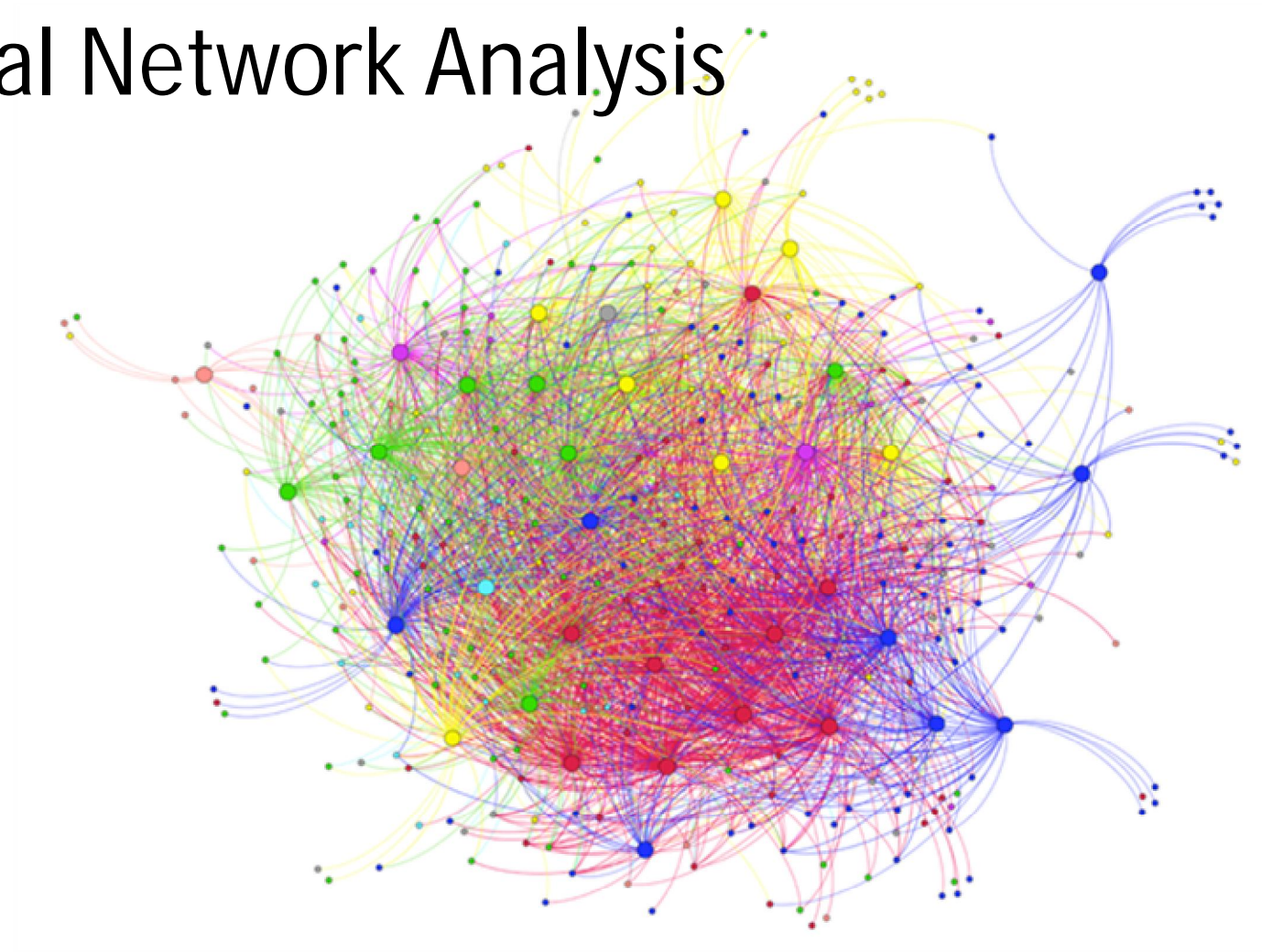
Selected chunk:  
Begin time: 00:00:18:707  
End time: 00:00:19:706  
Text: -



v  
vene  
ve  
ne  
v  
e  
r  
e  
Vene.  
Vene ono merel.  
Levo ono merel.  
Ivo ono Levon veli.  
Veli ono merel.  
Motori ono merel.

Links to pictures  
and manuscripts

# Social Network Analysis





# Web Tools

**KIELIPANKKI**  
The Language Bank of Finland

LANGUAGE BANK ACCESS CORPORA **TOOLS** FORUM ORGANIZATION SUPPORT SUOMEKSI PÄ SVENSKA

## Corpus interfaces



Korp is a web-based concordance tool that can be used for text corpus queries based on morphosyntactic analysis.



LAT (Language Archive Tools) is a tool for browsing, querying and sharing annotated speech and video corpora.



Mylly – a versatile data analysis platform with interactive visualizations and workflows



Some corpora are available for download via Kielipankki Downloads.



WebAnno annotation tool



Signbank – a lexical database of Finnish sign language



The Bank of Finnish Terminology in Arts and Sciences is a multidisciplinary project which aims to gather a permanent terminological database for all fields of research in Finland.



Query all CLARIN centers' corpora at once with CLARIN Federated Content Search.



Researcher of the Month: Markus Juutinen

### News

- Researcher of the Month: Markus Juutinen (10.10.2017)
- Network maintenance 10.10.2017 6:30am-9:00am (27.9.2017)
- Researcher of the Month: Laura-Maija Suur-Askola (1.9.2017)
- The Italian Letters from the Sixteenth Century Corpus in Kielipankki (30.8.2017)
- Language Bank newsletter 3.8.2017 (3.8.2017)

[More news](#)

### Tulevat tapahtumat

Verkkokurssi: Korpuslingvistiikka ja tilastolliset menetelmät 5.9.2017–20.10.2017

Kielipankki esittäytyy Historiantutkimuksen päivillä Turussa

# Training and User Involvement

## Training and Education

<http://clarin.eu/content/knowledge-centres>: the CLARIN Knowledge Sharing Infrastructures take care of the sharing of knowledge and expertise, education, training and dissemination

## Community Engagement

<https://www.clarin.eu/events>: CLARIN PLUS runs and regularly organizes expert seminars and workshops as well as researcher exchange programs

Past workshops:

- Exploring Historical Sources
- Exploring Oral History Archives 2016
- Working with Parliamentary Records
- Creation and Use of Social Media Resources
- Culture & Technology 2017



# FIN-CLARIN

## Training

- Basics of Text Analytics and Corpus Linguistics, 5 cr, (Information retrieval with Korp)
- Basics of Speech Annotation and Analysis, 5 cr, (Praat and ELAN)
- Corpus Clinic, 5 cr (Data management, Annotation methods and tools, RStudio)

## FIN-CLARIN roadshows and events

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

## Corpus and data-related advice:

[fin-clarin@helsinki.fi](mailto:fin-clarin@helsinki.fi)

## Technical support (servers, access rights, virtual workspace etc.):

[kielipankki@csc.fi](mailto:kielipankki@csc.fi)

# The Language Bank of Finland's Researchers of the Month

An archive of the previous months' Researcher of the Month interviews.

	2017	2016
1	Risto Turunen	
2	Jani Marjanen	
3	Tommi Jantunen	Päivi Pasanen
4	Jarmo Jantunen	Anna Dannenberg
5	Ilmari Ivaska	Marko Pantermöller
6	Juho Härme	Mihail Kopotev
7	Katja Västi	Kirsi-Maria Nummila
8	Paul-Thor Holmberg	Antti Kanner
9	Laura-Maija Suur-Askola	Tuija Määttä
10	Markus Juutinen	Auroora Vihervalli
11		Markus Hamunen
12		Hanna Westerlund



Researcher of the Month: Markus Juutinen

## News

- Researcher of the Month: Markus Juutinen (10.10.2017)
- Network maintenance 10.10.2017 6:30am-9:00am (27.9.2017)
- Researcher of the Month: Laura-Maija Suur-Askola (1.9.2017)
- The Italian Letters from the Sixteenth Century Corpus in Kielipankki (30.8.2017)
- Language Bank newsletter 3.8.2017 (3.8.2017)

[More news](#)