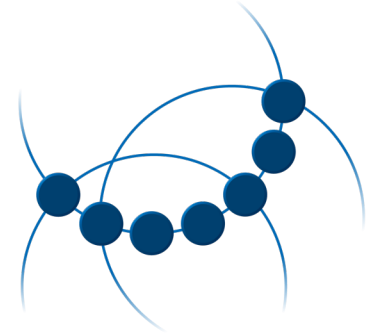




CLARIN
Common Language Resources and
Technology Infrastructure



FIN-CLARIN and CLARIN

Digital Research Infrastructure for the Humanities and Social Sciences

Mietta Lennes

Project Planning Officer, FIN-CLARIN

fin-clarin@helsinki.fi



CLARIN ERIC

European Research Infrastructure Consortium
founded on February 29, 2012

<https://www.clarin.eu> / NL



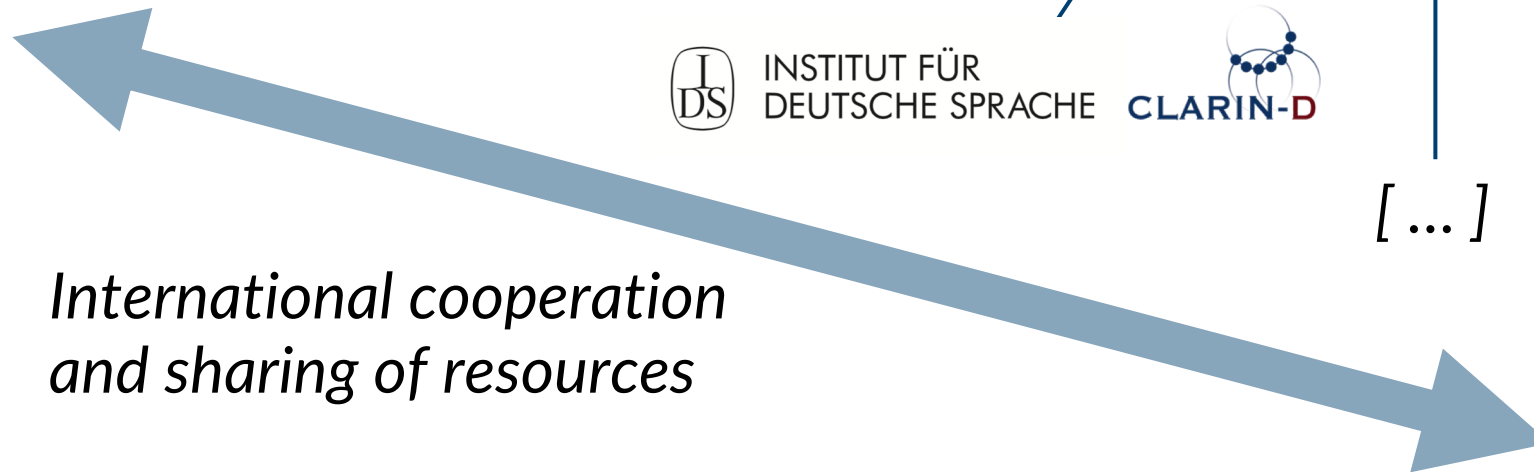
SWE-CLARIN



INSTITUT FÜR
DEUTSCHE SPRACHE



[...]



*International cooperation
and sharing of resources*

Member countries:

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



Member countries:

The Netherlands

Austria

Bulgaria

Czech Republic

Denmark

DLU

Estonia

Finland

Germany

Greece

Hungary

Italy

Latvia

Lithuania

Norway

Poland

Portugal

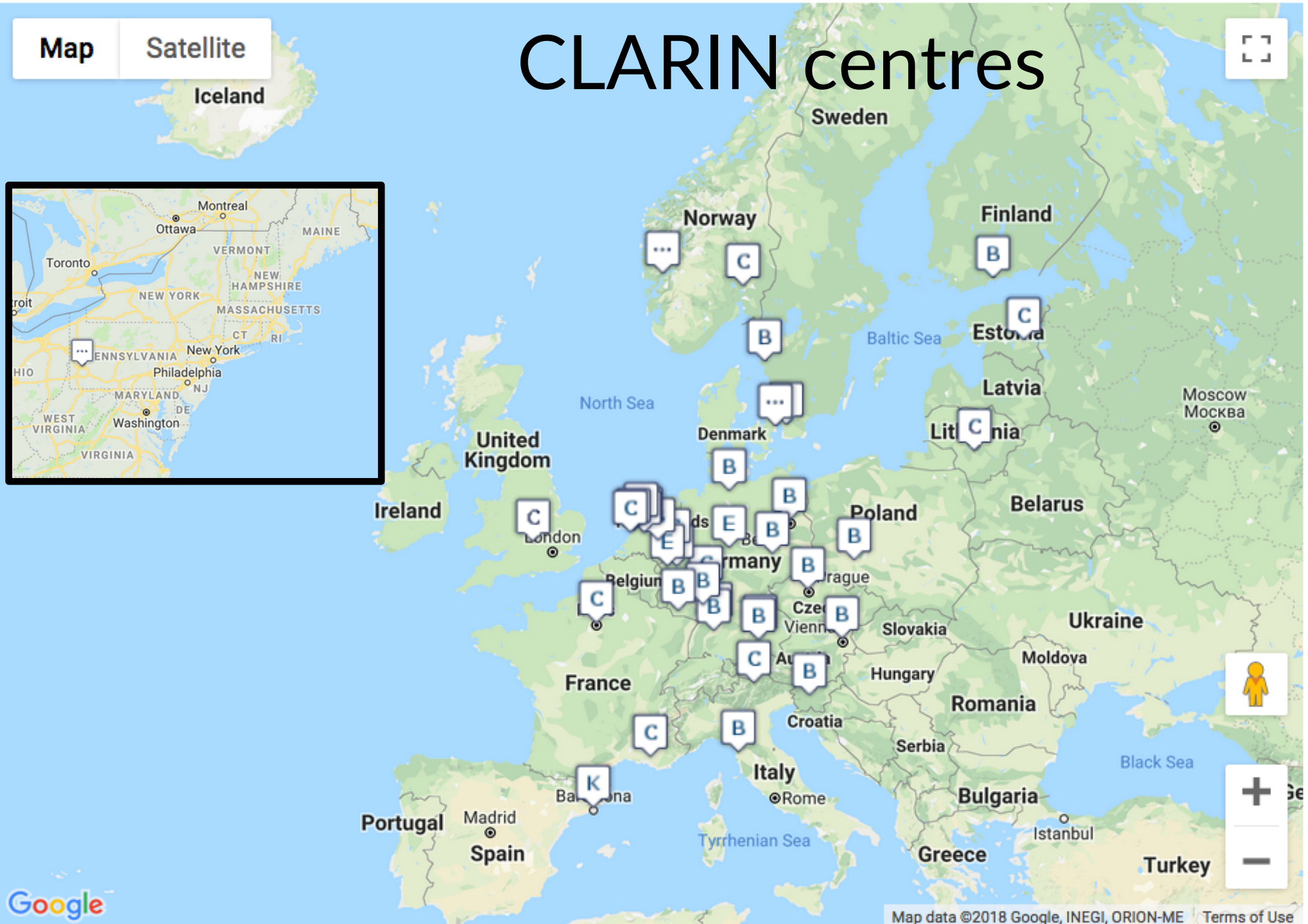
Slovenia

Sweden

France

UK

USA / CMU



FIN-CLARIN Partners

- University of Helsinki
- CSC – IT Center for Science



Coordination and access to large **centrally acquired resources 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



Access to **resources and tools developed locally** by individual researchers or research groups

CLARIN



Common Language Resources and Technology Infrastructure

FIN-CLARIN Partners

- 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

Coordination and access to large **centrally acquired resources and tools**

Access to **resources and tools developed locally** by individual researchers or research groups



CLARIN



Common Language Resources and Technology Infrastructure

https://vlo.clarin.eu

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



Showing all 1640603 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; Englisch 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...



https://vlo.clarin.eu

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

about
1,6 million
records

Search

Showing all 1640603 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; Englisch 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...



CLARIN Virtual Language Observatory

Welcome to the VLO!

Use the **search bar** below to search and browse everything and use **filters** to narrow down your results.

[See all records](#)

[Take a quick tour](#)

Search

Showing all 1640603 records

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

Language

Collection

Resource type

Modality

Format

Keyword



CLARIN RESOURCE FAMILIES

- Computer-mediated communication corpora
- Historical corpora
- L2 learner corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- ...

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...

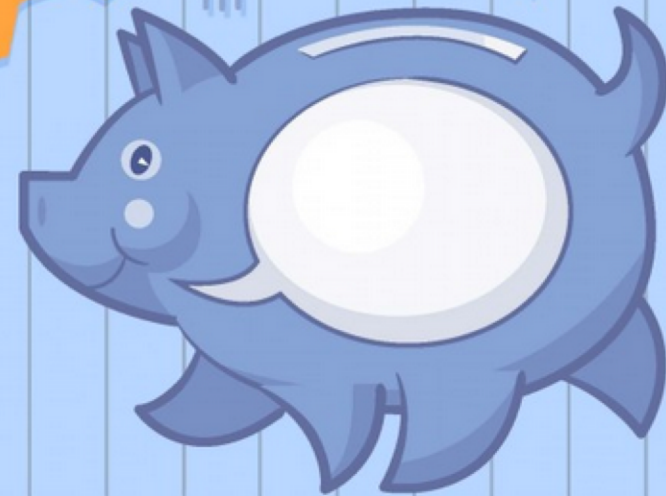
Resources	2018	2022
Text		
Magazines and newspapers 1770- (NLF and Web publ.)	12 Gw	20 Gw
Social media and similar sources 2000- (Suomi24, Ylilauta, ...)	4 Gw	10 Gw
Literature and manuscripts (Gutenberg, Fennica, archives)	60 Mw	70 Mw
Speech		
News broadcasts (YLE)		10000 h
Video sessions from the Finnish 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

Currently, FIN-CLARIN has approx. 19 GW in >1400 databases

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












KIELIPANKKI

The Language Bank of Finland



LANGUAGE BANK ACCESS CORPORA TOOLS ORGANIZATION SUPPORT

SUOMEKSI PÅ SVENSKA

Tool	Metadata	Manual	Admin
 KORP Korp Web-based concordance tool that can be used for text corpus queries based on morphosyntactic analysis.			
 LAT (Language Archive Tools) Tool for browsing, querying and sharing annotated speech and video corpora.			
Mylly Mylly Versatile data analysis platform with interactive visualizations and workflows.			
 Download Download certain corpora.			

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

KIELIPANKKI












The Language Bank of Finland



15 tools
currently listed

LANGUAGE BANK ACCESS CORPORA TOOLS

SUOMEKSI PÅ SVENSKA

Tool	Metadata	Manual	Admin
 KORP Korp Web-based concordance tool that can be used for text corpus queries based on morphosyntactic analysis.			
 LAT (Language Archive Tools) Tool for browsing, querying and sharing annotated speech and video corpora.			
Mylly Mylly Versatile data analysis platform with interactive visualizations and workflows.			
 Download Download certain corpora.			



Search & visualize text, speech and lexical data

Yksinkertainen Laajennettu Edistynyt Vertailu

Mannerheim (substantii) Etsi

myös alkuosa loppuosa ja samaista plen- ja suuraakoset

virkkeistä jotka sisältävät

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

Konkordanssi Tilastoja Sanakuva Kartta Nimiluokittelu

Paikkoja: 172

SUOMI24 2001–2014 (NÄYTE)

...tössä kranaatinheitimiä 81 Krh/32. Mutta lahtari luuseri Mannerheim huusi puhelimesta ja käski ampua kranaatteja
...oli käytössä kranaatinheitimiä 81 Krh/32. Fasisti Mannerheim huusi alaisille puhelimesta ja käski ampua kranaatteja
.../wiki/81_Krh/32 Lukijoiden on nykyään hyvä huomata, että Kremlin väki pitää yhä Mannerheimia ja Hitleriä syyllisinä
...vuonna 1971 kuolin vuoteella oleva mies kertoi että Mannerheim juovuksissa antoi käskyn ampua
...Jos ajattelemme
...Mainilan kylään eli aloittaa laitton sota Neuvostoliittoa vastaan.
...Mainilan kylään.
...Mainilan laukauksiin.
...Mainilaa .Toinen merkittävä tapaus näiltä ajoilta on Knut Bösse Viipurin rannikkovartioston päällikkö, joka harjoitteli van
...Mainilan laukauksia, miehisen tajunnan omistanut Mannerheim teki silloin todennäköisesti sotarikoksen.
...Mainila ei tuolloin ollut Suomen tykistön kantaman sisällä, koska tykit oli Mannerheimin nimeno maisesta käskystä ve
...1) Mainilassa lahtari Mannerheim huusi puhelimesta ja käski ampua kranaatinheitimillä 81 Krh/32 Mainilan kylään.
...Mainilan kylään.
...Mainilan kylään.
...Mainilan laukausten ammutti Zhdanov joka myöhemmin kyykytti Mannerheimia otti häneltä matkustusasiakirjat pois j
...Mainilan laukausten ammuttaja Zhdanov yritti rangaista Mannerheimia mutta Stalin lähetti oikean kätensä Helsinkiin p

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

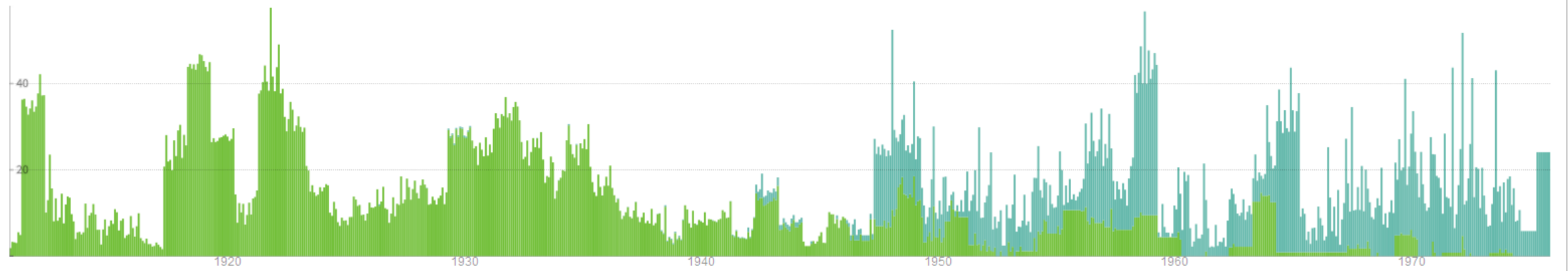
Konkordanssi Tilastoja Sanakuva Kartta Nimiluokittelu Kuvaaja

Viiva Pylväs Taulukko

Aikatieto puuttuu 0.03%:sta valitusta aineistosta

Ajanjaksolta ei ole tietoja

- ✓ sosiaalihuolto
- ✓ köyhäinhoito



Datasets

- Datasets
 - query.cqp.txt
 - query.cqp-s3527p0.korp.json
 - query.cqp-s3527p0.korp-tokens.tsv
 - query.cqp-s3527p0.korp-meta.tsv

Analysis tools

Kielipankki

- Korp search
- Korp data
- Relation algebra
- Simple statistics
- Advanced statistics
- Export formats
- Syntactic analysis
- Morphological analysis
- Speech recognition
- Preprocessing
- Finite-State Technology
- Finite-State Transducers
- Job management

KWIC as TSV

- KWIC 2-grams in TSV
- KWIC 3-grams in TSV
- KWIC dependency triples in TSV

 Show parameters
Run 

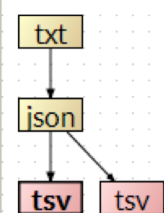
Korp JSON-form concordance as two TSV files, tokens with their annotations in one and structural annotations in the other. Both files contain a sentence counter attribute so that they can be easily joined into one.

Mylly – “The Mill”

More help

Show tool sourcecode

Workflow

 Fit


Visualisation

Spreadsheet

 Maximise

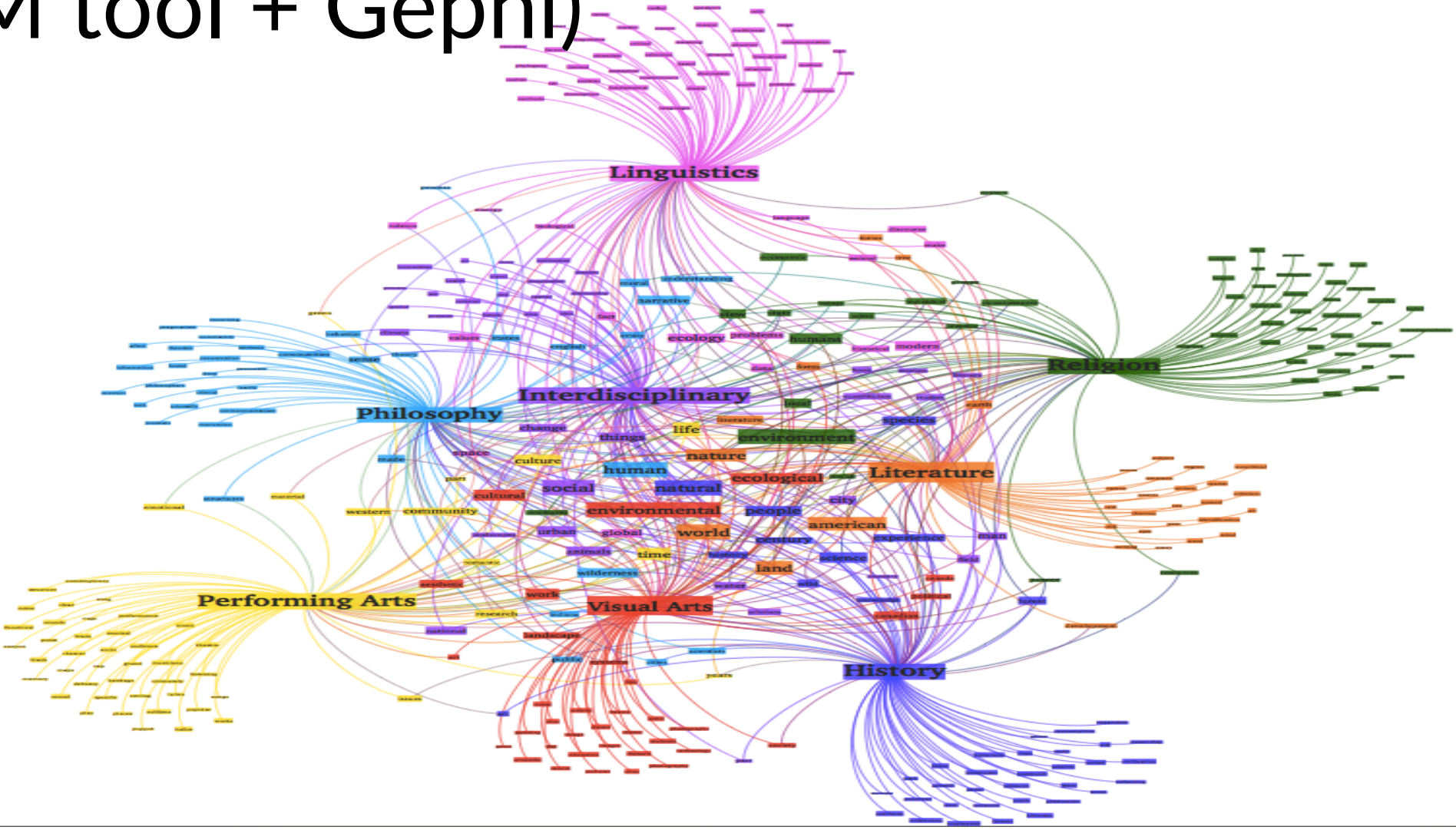
 Detach

 Close

Showing 3609 rows of 3609 and all 12 columns

kMmatch	kMsen	kMtok	lemma	deprel	ref	aid	pos	nertag	word	msd	deprel
0	0	0	silloin	advmod	1	a12594	Adv	-	Silloin	CASECHA...	3
0	0	1	täytyä	aux	2	a12595	V	-	täytyy	PRS_Sg3 V...	3
0	0	2	uskaltaa	ROOT	3	a12596	V	-	uskaltaa	NUM_Sg C...	0
0	0	3	nostaa	xcomp	4	a12597	V	-	nostaa	NUM_Sg C...	3
1	0	4	kissa	dobj	5	a12598	N	-	kissa	NUM_Sg C...	4
0	0	5	pöytä	nommod	6	a12599	N	-	pöydälle	NUM_Sg C...	4
0	0	6	arvoisa	amod	7	a12600	A	-	arvoisa	NUM_Sg C...	8
0	0	7	puhe mies	dobj	8	a12601	N	-	puhemies	NUM_Sg C...	4
0	0	8	.	punct	9	a12602	Punct	-	.	-	3
0	1	0	ja	advmod	1	a17480	Adv	-	Ja	CASECHA...	5
0	1	1	sitten	advmod	2	a17481	Adv	-	sitten	-	3
0	1	2	vihonviimei...	acomp	3	a17482	A	-	vihonviimei...	NUM_Sg C...	5
0	1	3	lopuksi	advmod	4	a17483	Adv	-	lopuksi	-	5
0	1	4	haluta	ROOT	5	a17484	V	-	haluan	PRS_Sg1 V...	0
0	1	5	nostaa	xcomp	6	a17485	V	-	nostaa	NUM_Sg C...	5
0	1	6	se	dobj	7	a17486	Pron	-	sen	SUBCAT_D...	6
1	1	7	kissa	poss	8	a17487	N	-	kissan	NUM_Sg C...	9
0	1	8	pöytä	nommod	9	a17488	N	-	pöydälle	NUM_Sg C...	6
0	1	9	joka	rel	10	a17489	Pron	-	jota	SUBCAT_R...	12
0	1	10	tämä	nommod	11	a17490	Pron	-	tässä	SUBCAT_D...	12
0	1	11	kiertää	rcmod	12	a17491	V	-	kiertetään	PRS_Pe4 V...	9
0	1	12	kuin	comparator	13	a17492	C	-	kuin	SUBCAT_CS	14

Topic Modelling (TM tool + Gephi)



CLARIN User Involvement

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

Workshops:

- [Exploring Spoken Word Data in Oral History Archives](#), 18-19 April 2016, Oxford (UK)
- [Working with Digital Collections of Newspapers](#), 19-21 September 2016, Leuven
- [Working with Parliamentary Records](#), 27-29 March 2017, Sofia (Bulgaria)
- [Creation and Use of Social Media Resources](#), 18-19 May 2017, Kaunas (Lithuania)
- [Workshop on interoperability of L2 resources and tools](#), 6-8 December, Gothenburg
- [CLARIN Workshop on Translation memories, corpora, termbases: Bridges between translation studies and research infrastructures](#), 8-9 February 2018, Vienna (Austria)
- [Parliamentary Records \(ParlaCLARIN@LREC2018\)](#), 7 May 2018, Miyazaki (Japan)

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

- **Kielipankki 20th Jubilee Roadshow events in FIN-CLARIN member organizations**
- **Speech annotation workshop / University of Turku**
- **Presentations at XLIV Finnish Conference of Linguistics / University of Jyväskylä**
- **Demo and Presentation at Language Center Days / University of Eastern Finland**
- **Demo and Presentation for Principal Investigators / UHEL Arts and Humanities**
- **CLARIN PLUS Workshop on User Involvement / University of Helsinki**
- **Demos at Historical Network Research conference / University of Turku**
- **Presentations in ComHIS Roadshow events**
- **Digital Humanities Hackathon (in collaboration with HELDIG)**
- **International conference series RDHum – Research Data and Humanities (to begin in Oulu, 14-16 August 2019)**

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)

Waveform and spectrogram analysis of a speech segment. The spectrogram shows frequency components over time. Below the spectrogram is a phonetic transcription table:

Ku	ma	(u)un	el	le	hän	si	?
ku	ma	uun	el	le	han	si	?



Corpus Clinic (5 cr)

Corpus Clinic interface showing a dependency tree for the sentence "Tänne sun huani no niin varmaai pit sä neuvoos isään maholl: itä mä abianinkatu e: ss sä tuut ni: i okei mä ooi lvä". The interface includes a bar chart showing the distribution of dependency relations and a spectrogram of the audio input.



Sharing expertise in online teaching





LANGUAGE BANK ACCESS CORPORA TOOLS ORGANIZATION SUPPORT

SUOMEKSI PÅ SVENSKA

The Language Bank of Finland's Researchers of the Month

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

	2018	2017	2016
1	Krista Lagus	Risto Turunen	
2	Matias Tamminen	Jani Marjanen	
3	Zsuzsi Máthé	Tommi Jantunen	Päivi Pasanen
4	Mari Siiroinen	Jarmo Jantunen	Anna Dannenberg
5	Ivana Kováčová	Ilmari Ivaska	Marko Pantermöller
6	Maximilian Murmann	Juho Härme	Mihail Kopotev
7	Mietta Lennes	Katja Västi	Kirsi-Maria Nummila
8	Maria Huttu-Hiltunen	Paul-Thor Holmberg	Antti Kanner
9		Laura-Maija Suur-Askola	Tuija Määttä
10		Markus Juutinen	Auroora Vihervalli
11		Eero Voutilainen	Markus Hamunen
12		Heli Tissari	Hanna Westerlund

Search the Language Bank Portal:

Haku ... Hae



Researcher of the Month: Annika Liiti

News

- Researcher of the Month: Annika Liiti (13.9.2018)
- Researcher of the Month: Maria Huttu-Hiltunen (6.8.2018)
- Researcher of the Month: Mieta Lennes (9.7.2018)
- Researcher of the Month: Maximilian Murmann (11.6.2018)
- Researcher of the Month: Ivana Kováčová (7.5.2018)

[More news](#)

National co-operation agreement between the ERICs of Digital Humanities



ESS ERIC
(University of Turku)



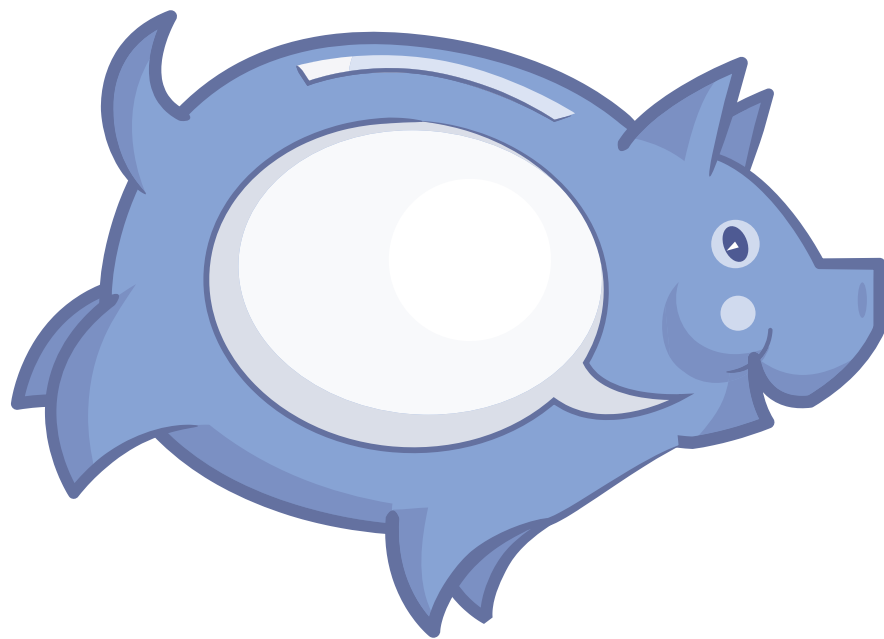
CESSDA ERIC
(FSD)



SHARE ERIC
(Family Federation
of Finland)



CLARIN ERIC
(FIN-CLARIN)



Kiitos! Thank you!

www.kielipankki.fi

General support

fin-clarin@helsinki.fi

Technical support

kielipankki@csc.fi