FIN-CLARIAH MEETING

December 1, Linna Building, Kalevantie 5, Tampere

PRELIMINARY PROGRAM

11.00-11.10 Welcoming Words by Sanna Kumpulainen, Associate Professor in Information Studies, Tampere University

11.10-12.00 Keynote I on Studying SSH Research Needs: Elina Late, Senior Research Fellow in Information Studies, Tampere University

12.00-13.00 Lunch

13.00-13.45 Keynote II on Language Models: Sampo Pyysalo, Associate Professor at the Department of Computing, University of Turku

13.45-14.30 Work Package Presentations I

13.45-14.00 WP1.3 ~~Veronika Laippala~~: Noise-Tolerant NLP

14.00-14.20 WP 1.1, 1.2, 2.1, 2.2, 2.3 Mietta Lennes: Kielipankki – The Language Bank of Finland

14.20-14.25 WP2.4 Harri Kettunen: Helsinki Term Bank for the Arts and Sciences

14.25-14.30 WP2.5 Jenny Tarvainen: Automated Text Tools for Learner Language

14.30-15.00 Coffee

15.00-16.00 Work Package Presentations II

15.00-15.05 WP3.1 Martin Matthiesen: Pipeline from the National Library to CSC

15.05-15.10 WP3.2 Tanja Välisalo: Named Entity Recognition for NARC Data

15.10-15.15 WP4.3 Eetu Mäkelä: Evaluation and Subsetting

15.15-15.20 W4.1 Julia Matveeva: Metadata Harmonization

15.20-15.25 WP4.4 Mikko Laitinen: Twitter

15.25-15.30 WP4.2 Eero Hyvönen: LOD

15.30-15.35 WP3.4 Raine Koskimaa: Game Streams

15.35-15.40 WP3.3 Maria Valaste: Qualitative Surveys

15.40-15.45 WP3.5 Kimmo Elo (Risto Turunen replacing): Text Networks

15.45-15.50 WP5 Sanna Kumpulainen: Evidence-based RI Development + Education & Resources

16.00-17.00 Free Chilling & Refreshments / Parallel session: Executive Board Meeting (with Zoom option)

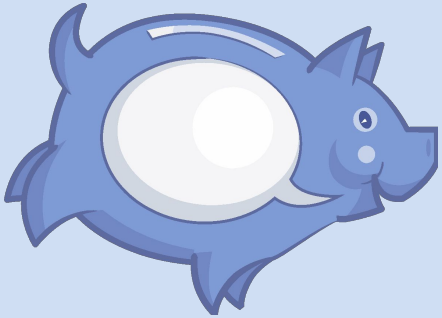
MATERIALS

The presentation slides will be made available after the event at
https://www.kielipankki.fi/organization/fin-clariah/fin-clariah-2023-12-01/

PROJECT DELIVERABLES

https://www.kielipankki.fi/organization/fin-clariah/deliverables/

WP1.3 NOISE-TOLERANT NLP, 15 MIN



KIELIPANKKI
The Language Bank of Finland

FIN-CLARIN

FIN-CLARIAH 2022-23:

Overview of the work completed in Modules 1 and 2

University of Tampere, 1.12.2023

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.

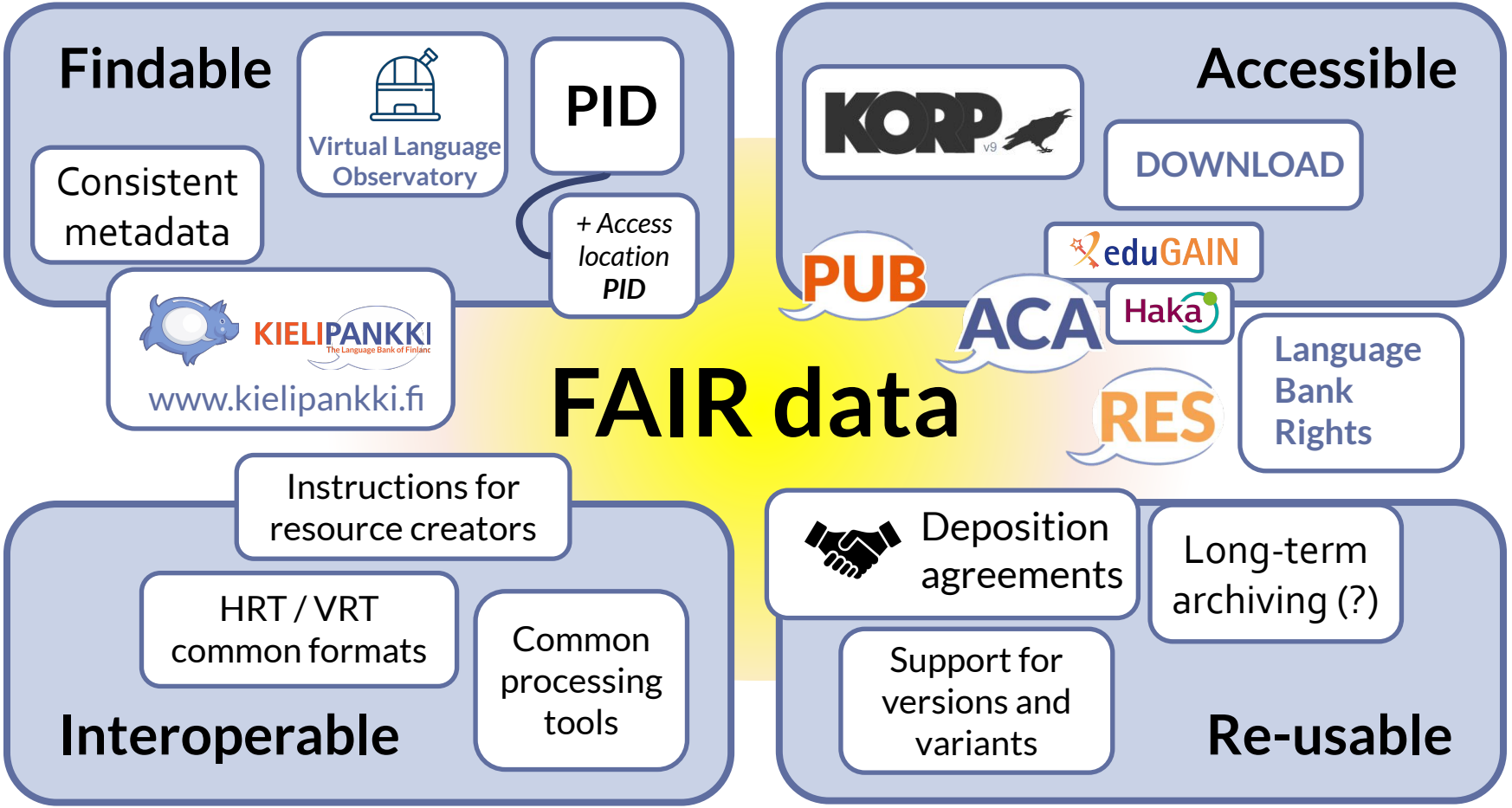
Search the Language Bank Portal:

Haku ... Hae

Researcher of the Month: Aleksi Sahala

News

- New resource (beta): Christmas Gospel text-to-speech in four Uralic languages, Korp (29.11.2023)


Findable
Virtual Language Observatory
PID
Consistent metadata
+ Access location PID
KIELIPANKKI
The Language Bank of Finland
www.kielipankki.fi
Accessible
KORP v9
DOWNLOAD
PUB
eduGAIN
ACA
Haka
Language Bank Rights
RES
FAIR data
Instructions for resource creators
HRT / VRT common formats
Common processing tools
Interoperable
Deposition agreements
Long-term archiving (?)
Support for versions and variants
Re-usable

Books and manuscripts

Parallel corpora and other multilingual resources

Newspapers and journals

Lexicons and terminologies

Sessions of the Parliament (videos + transcripts)

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

Learner language corpora

Sign languages

Lahjoita puhetta – Donate Speech (puhelahjat)

Dialect corpora

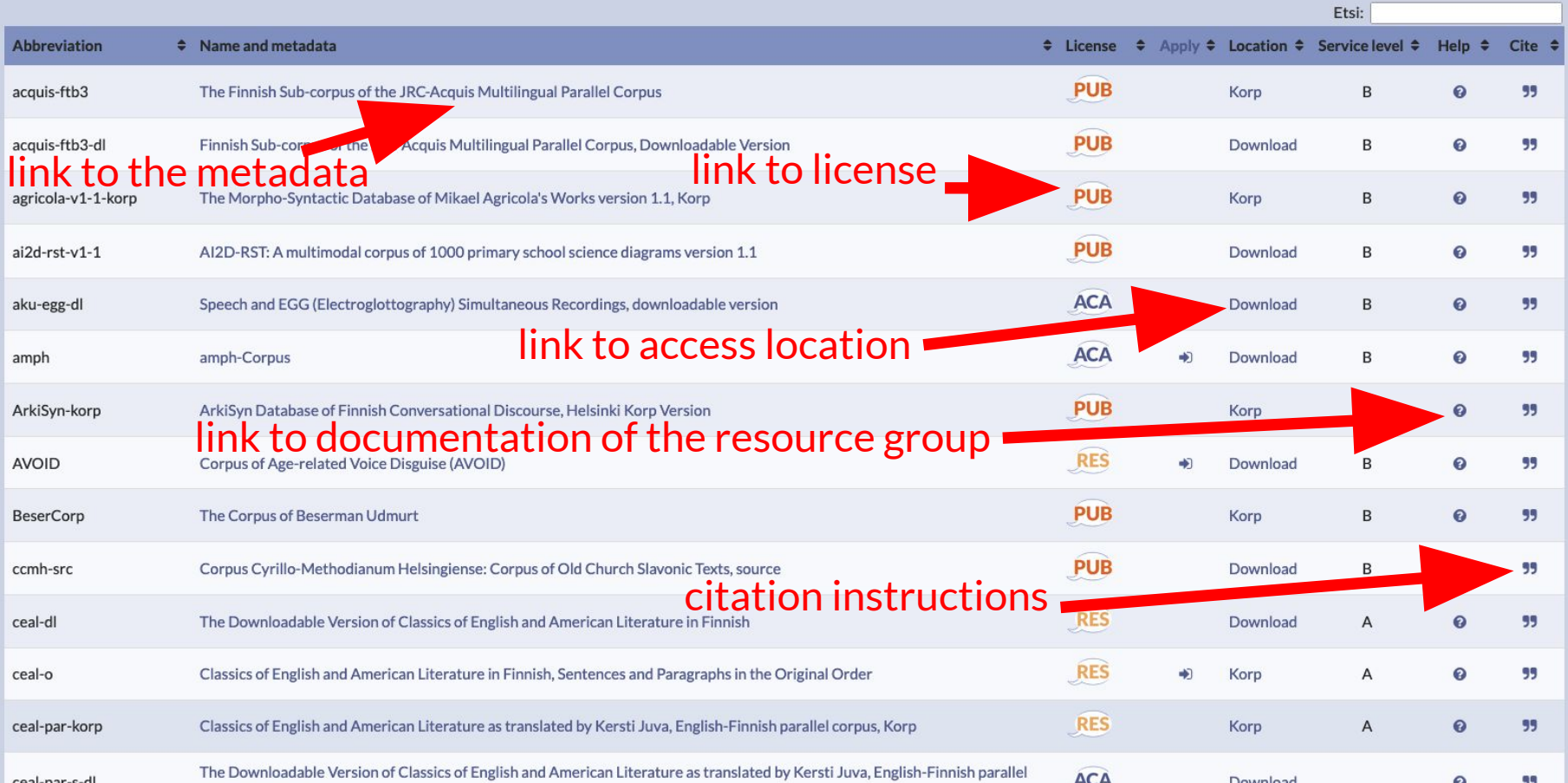
See the complete list of corpora at https://www.kielipankki.fi/corpora/

https://www.kielipankki.fi/corpora

Etsi:

Abbreviation	Name and metadata	License	Apply	Location	Service level	Help	Cite
acquis-ftb3	The Finnish Sub-corpus of the JRC-Acquis Multilingual Parallel Corpus	PUB		Korp	B		
acquis-ftb3-dl	Finnish Sub-corpus of the JRC-Acquis Multilingual Parallel Corpus, Downloadable Version	PUB		Download	B		
agricola-v1-1-korp	The Morpho-Syntactic Database of Mikael Agricola's Works version 1.1, Korp	PUB		Korp	B		
ai2d-rst-v1-1	AI2D-RST: A multimodal corpus of 1000 primary school science diagrams version 1.1	PUB		Download	B		
aku-egg-dl	Speech and EGG (Electroglottography) Simultaneous Recordings, downloadable version	ACA		Download	B		
amph	amph-Corpus	ACA		Download	B		
ArkiSyn-korp	ArkiSyn Database of Finnish Conversational Discourse, Helsinki Korp Version	PUB		Korp	B		
AVOID	Corpus of Age-related Voice Disguise (AVOID)	RES		Download	B		
BeserCorp	The Corpus of Beserman Udmurt	PUB		Korp	B		
ccmh-src	Corpus Cyrillo-Methodianum Helsingiense: Corpus of Old Church Slavonic Texts, source	PUB		Download	B		
ceal-dl	The Downloadable Version of Classics of English and American Literature in Finnish	RES		Download	A		
ceal-o	Classics of English and American Literature in Finnish, Sentences and Paragraphs in the Original Order	RES		Korp	A		
ceal-par-korp	Classics of English and American Literature as translated by Kersti Juva, English-Finnish parallel corpus, Korp	RES		Korp	A		
ceal-par-s-dl	The Downloadable Version of Classics of English and American Literature as translated by Kersti Juva, English-Finnish parallel	ACA		Download			

https://www.kielipankki.fi/corpora



Citation instructions

amph	amph-Corpus	ACA		Download	B		
ArkiSyn-korp	ArkiSyn Database of Finnish Conversational Discourse, Helsinki Korp Version	PUB		Korp	B		
AVOID	Corpus of Age-related Voice Disguise (AVOID)	RES		Download	B		
BeserCorp	The Corpus of Beserman Udmurt	PUB		Korp			
ccmh-src	Corpus Cyril… Slavonic Text…						

[suomeksi] [in English]

Reference instructions: AVOID

Please cite the language resource as follows:

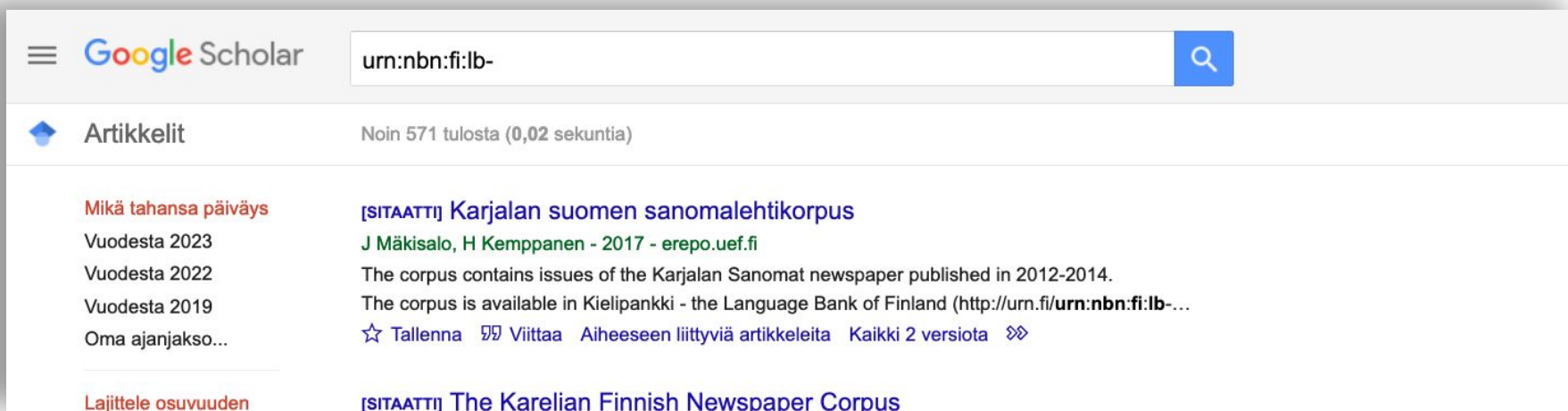
Kinnunen, T., Hautamäki, R. G., Sahidullah, M., Hautamäki, V., Werner, S., & Bentz, M.. *Corpus of Age-related Voice Disguise (AVOID)* [speech corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2018060621

Show: [Bibtex] [Zotero]

Search for references to the language resource in Google Scholar

Do researchers follow the citation instructions for corpora & resources?

- Google Scholar 2023-11-30:
  571 references cite the URN prefix of the Language Bank of Finland (*urn:nbn:fi:lb-*)



Platforms and services provided by CSC

META-SHARE LO:
curated metadata catalogues for resources

Download service:
downloadable versions of resources

KORP
concordances and statistics from corpora; links to external resources

Demo tools:
examples of software for text and speech analysis & processing

HeLI-OTS:
automatic language identification from text

Aalto-ASR, Tekstiks:
automatic speech recognition (ASR) tools

>> Language models…

See the complete list of tools at https://www.kielipankki.fi/tools/

The Language Bank of Finland Researchers of the Month

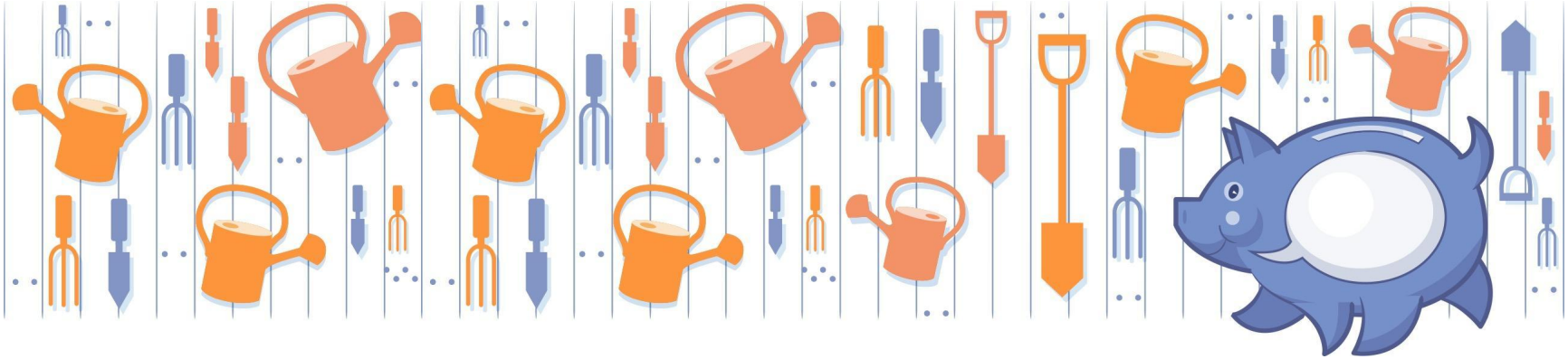
Suomeksi

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
2023.06	Mikael Varjo	ArkiSyn	2022, 2020, 2019, 2018
2023.05	Rosa González Hautamäki	AVOID	2019, 2018, 2017, 2016
2023.04	Johanna Vaattovaara	Suomi24	2022a, 2022b, 2019, 2012, 2011, 2009
2023.03	Noora Hoffrén	CFINSL, snowfrog-dl	2019
2023.02	Maria Sarhemaa	Suomi24	2022, 2021

KIELIPANKKI
The Language Bank of Finland

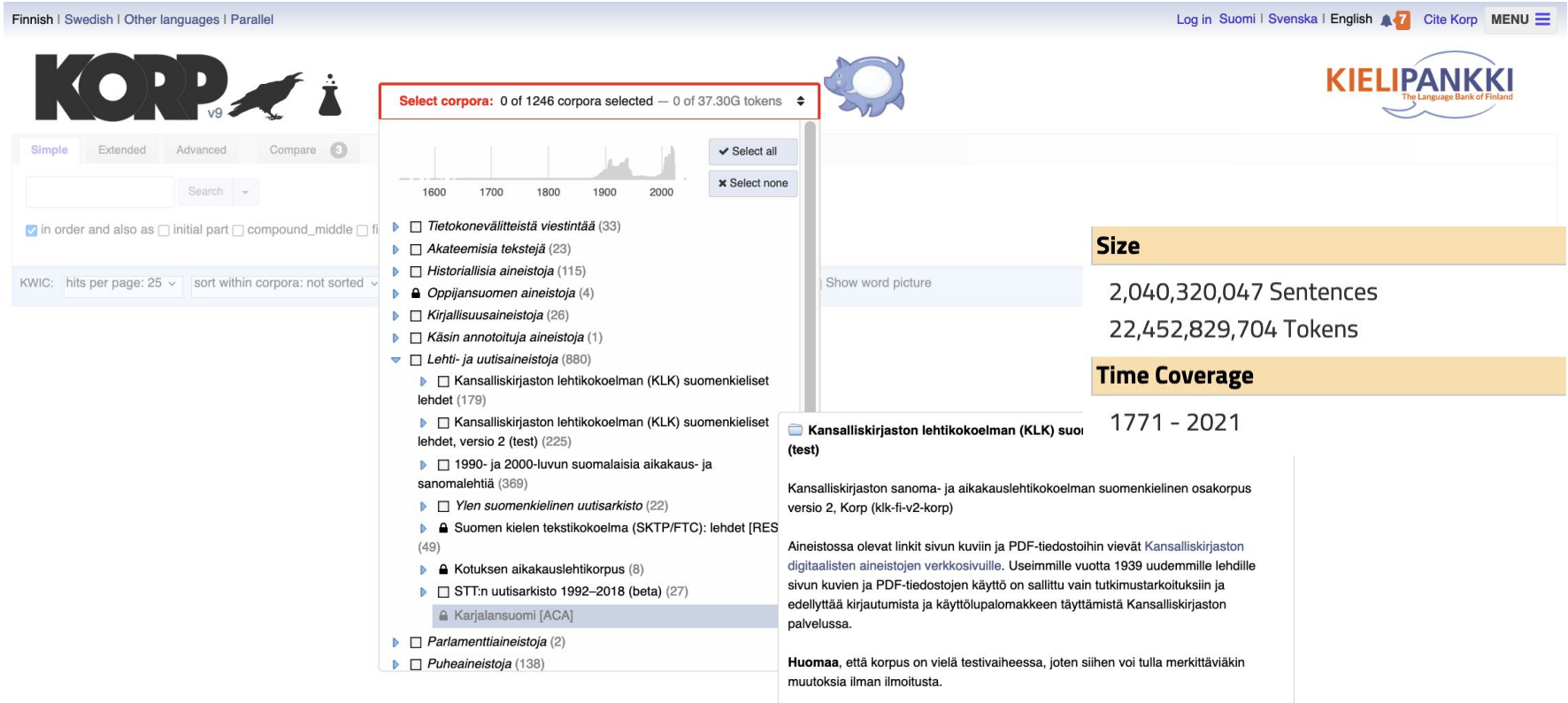


WP1.1:

TEXT PROCESSING AND ANNOTATION ENVIRONMENTS

FIN-CLARIN

The Finnish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland version 2, Korp (*klk-fi-v2*) http://urn.fi/urn:nbn:fi:lb-202009152



In 2020 and 2021, get statistics by the name of publication (search for 'any word')

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

KWIC | Statistics | Word picture

Statistics by the name of publication (search for ’any word’) in 2020-2021 (*klk-fi-v2*)

Show Trend Diagram | Show map

Number of rows: 301

publication	Total	KLK suomi v2: 2020	KLK suomi v2: 2021
Σ	1 000 000 (900 280 994)	1 000 000 (537 488 468)	1 000 000 (362 792 526)
Helsingin Sanomat	60 091 (54 098 749)	53 738,8 (28 883 982)	69 501,9 (25 214 767)
Kaleva	30 475,7 (27 436 662)	27 746,6 (14 913 470)	34 518,9 (12 523 192)
Aamulehti	30 133 (27 128 185)	31 520,6 (16 941 960)	28 077,3 (10 186 225)
Ilta-Sanomat	27 548,9 (24 801 789)	24 395,1 (13 112 080)	32 221,5 (11 689 709)
Satakunnan Kansa	27 221,2 (24 506 771)	24 712,4 (13 282 606)	30 938,2 (11 224 165)
Turun Sanomat	26 823,8 (24 148 960)	26 775,4 (14 391 468)	26 895,5 (9 757 492)
Etelä-Suomen Sanomat	25 511,2 (22 967 280)	23 027,7 (12 377 137)	29 190,6 (10 590 143)
Lapin Kansa	24 648,8 (22 190 848)	23 039,7 (12 383 552)	27 032,8 (9 807 296)
Ilkka-Pohjalainen	24 100,2 (21 696 991)	21 497,8 (11 554 840)	27 955,8 (10 142 151)
Keskipohjanmaa	23 916,5 (21 531 589)	21 454,5 (11 531 548)	27 564,1 (10 000 041)
Savon Sanomat	22 006 (19 811 571)	19 544 (10 504 666)	25 653,5 (9 306 905)
Keskisuomalainen	20 734,3 (18 666 739)	21 396,5 (11 500 380)	19 753,3 (7 166 359)
Karjalainen	20 191,2 (18 177 773)	20 279 (10 899 721)	20 061,2 (7 278 052)
Hämeen Sanomat	19 833,2 (17 855 486)	18 257,4 (9 813 119)	22 168 (8 042 367)
Iltalehti	17 134 (15 425 423)	15 278,6 (8 212 076)	19 882,8 (7 213 347)
Länsi-Savo	16 873,6 (15 190 967)	14 872,8 (7 993 951)	19 837,8 (7 197 016)
Salon Seudun Sanomat	15 075,9 (13 572 560)	14 015,8 (7 533 315)	16 646,6 (6 039 245)

KORP v9

2 of 1246 corpora selected — 900.28M of 37.30G tokens

KIELIPANKKI The Language Bank of Finland

[word = "lap.*"]

Simple | Extended | Advanced | Compare 3

Filter: Add publication

base form | is
korona Aa
or
and | Options

Add token
Add boundary

Search | within sentence

Base form *korona* in 2020-2021 (*klk-fi-v2*)

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

KLK suomi v2: 2020: 133941

Results: 220,861

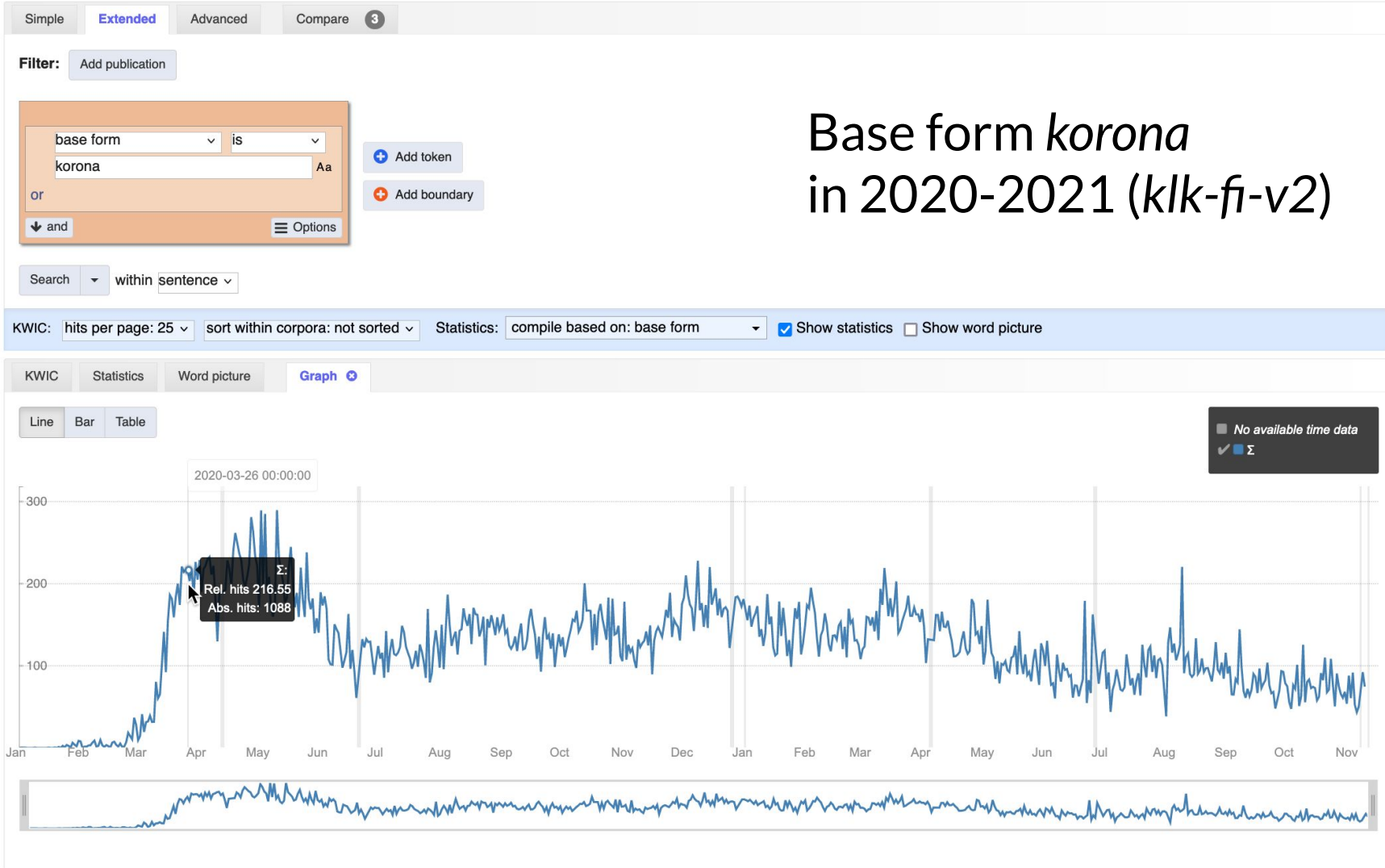
« ‹ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ... › » Go to page of 8835 Show context

	KLK SUOMI V2: 2021 (does not support extended context)	
	Koronan	takia hiihtokausi on rikkonainen.
Hän syytti laittomasta pidätyksestään ” bunkkeripap paa ”,	**koronan**	takia virkaasun nossaan pysyttelevää presidentti Vladimi
” He myös saavat heikompaa terveydenhuoltoa. ”	**Koronan**	seurauksena jo ennestään olemassa ollut rotulinjojen mu
Kuten sairastamaan	**koronan**	, äijäparka.
Ennuste ennen	**koronaa**	
IEA on mallintanut erilaisia skenaarioita, miten energian kysyntä kehittyy	**koronan**	jälkeisenä aikana.
	koronan	jälkiltä tarjonta onkin huonompaa, se vähentää joukkoliik

CORPUS
KLK suomi v2: 2021

Subcorpus of: Kansalliskirjaston lehtikokoelman (KLK) suomenkieliset lehdet, versio 2
Metadata
Cite corpus
Link to corpus in Korp
Persistent identifier: urn:nbn:fi:lb-202009152
Licence: CC BY 4.0 (CLARIN PUB)

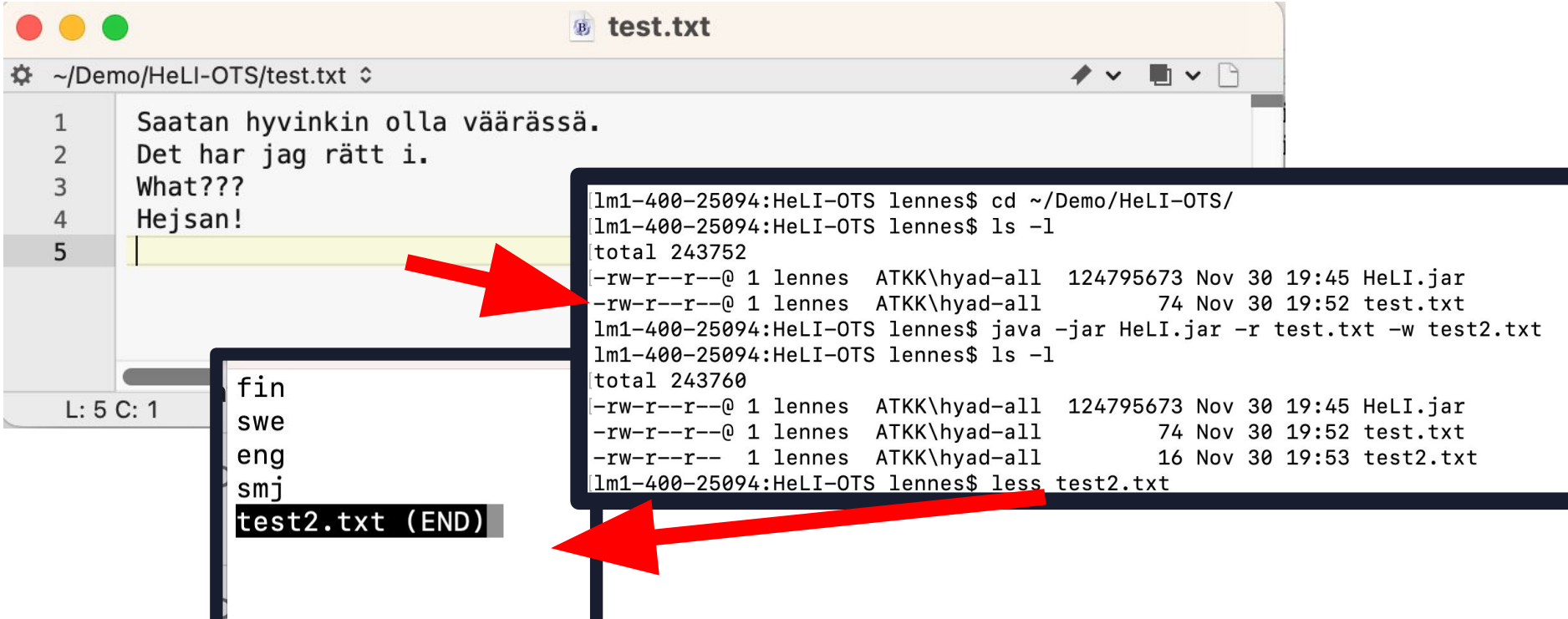
TEXT ATTRIBUTES



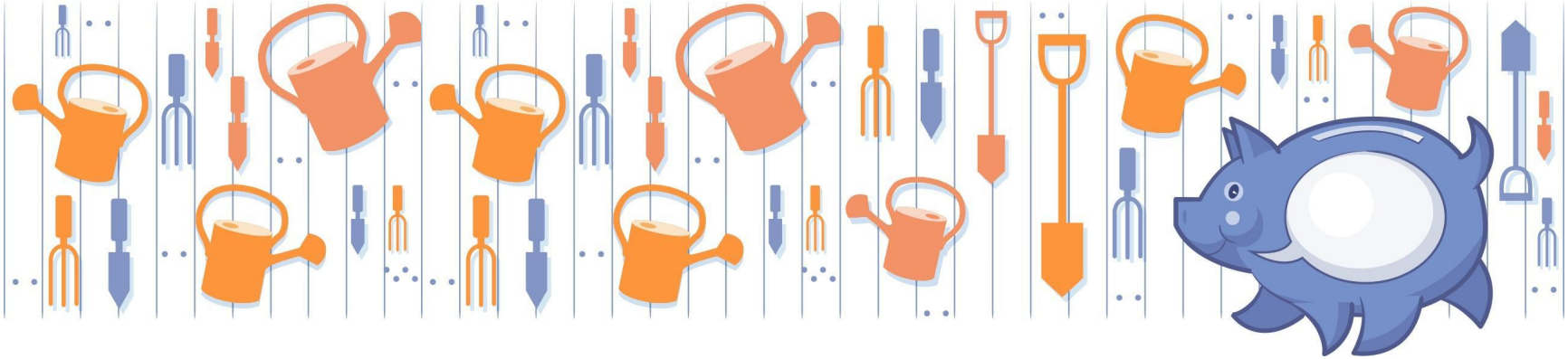
Base form *korona* in 2020-2021 (*klk-fi-v2*)

Language identification from text: HeLI-OTS 1.5

Available on Zenodo: https://doi.org/10.5281/zenodo.10071264



KIELIPANKKI
The Language Bank of Finland



WP1.2:

SPEECH PROCESSING AND ANNOTATION

FIN-CLARIN

Tekstiks

HOME FILES DEMO EN ML

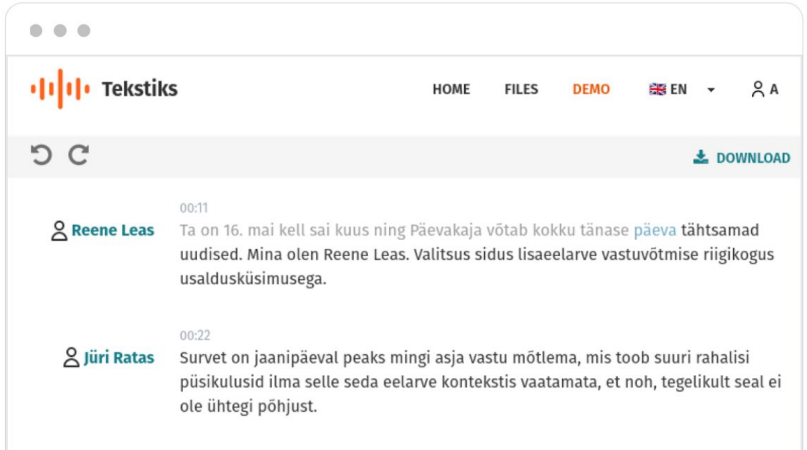
Speech recognition: speech to text

Automated speech transcription service for Estonian speech and a user interface for transcription editing.

GET STARTED VIEW DEMO

How does it work?

Tekstiks.eeis a public speech recognition service of TalTech's Laboratory of Language Technology. The system demonstrates technologies and models developed in this lab. These currently achieve state-of-the-art Estonian speech recognition results, even compared to commercial alternatives. The system is fully automated and can process multiple files in parallel. There can be a queue and delays, especially during business hours. The average processing time is about half of the recording's length. Coming soon: real-time progress and expected time of completion information



1. Upload a speech recording in Estonian

Most of the popular audio and video formats are supported. Max size limit is 500MB

2. Wait for the speech recognition to complete.

The system trained using machine learning methods will search for Estonian speech segments and tries to differentiate multiple speakers. Then it will transcribe the speech segments into text and finally will add punctuation. Many Estonain celebrities and radio personalities can be identified by name as well.

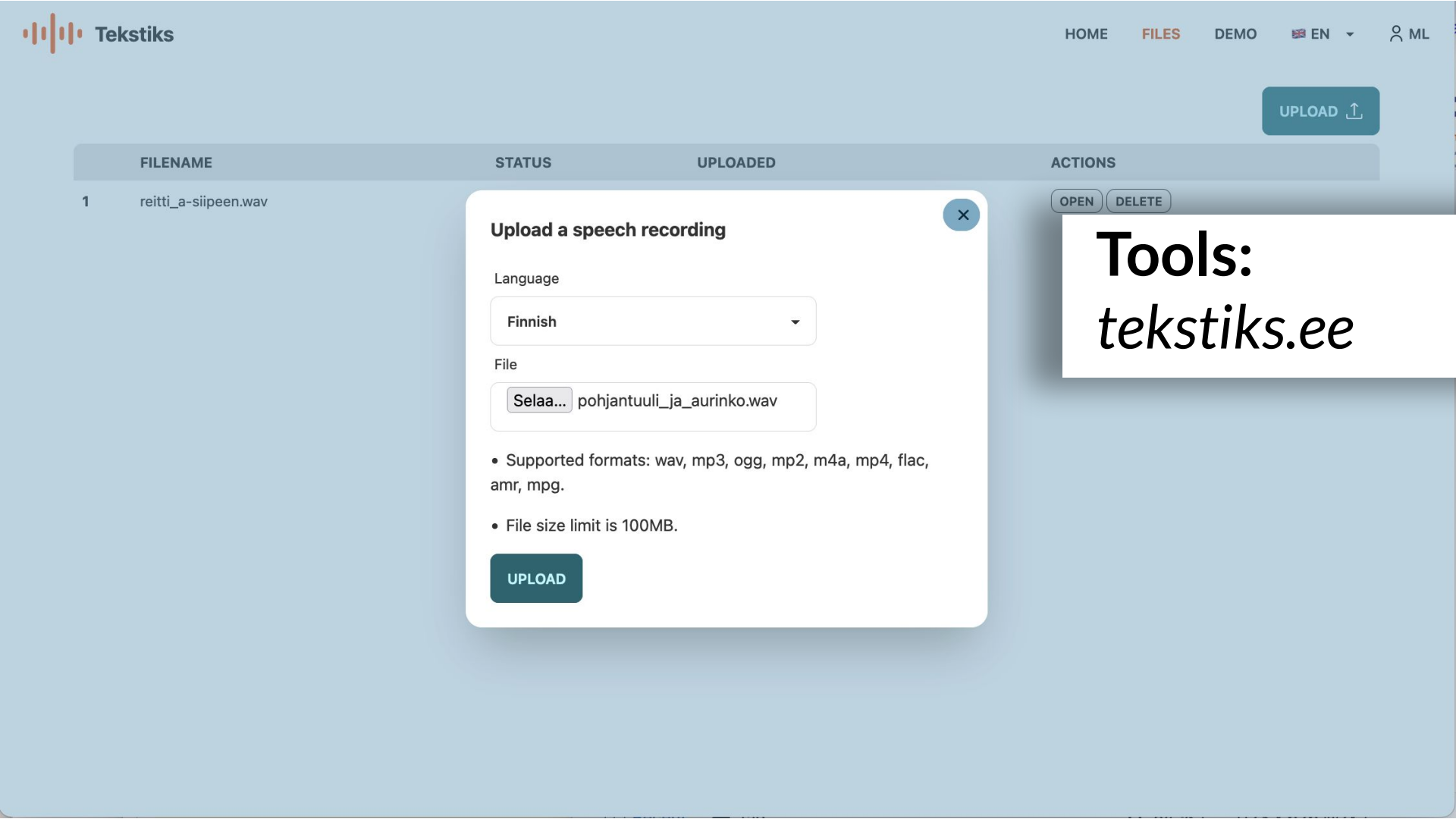
3. Correct speech recognition mistakes

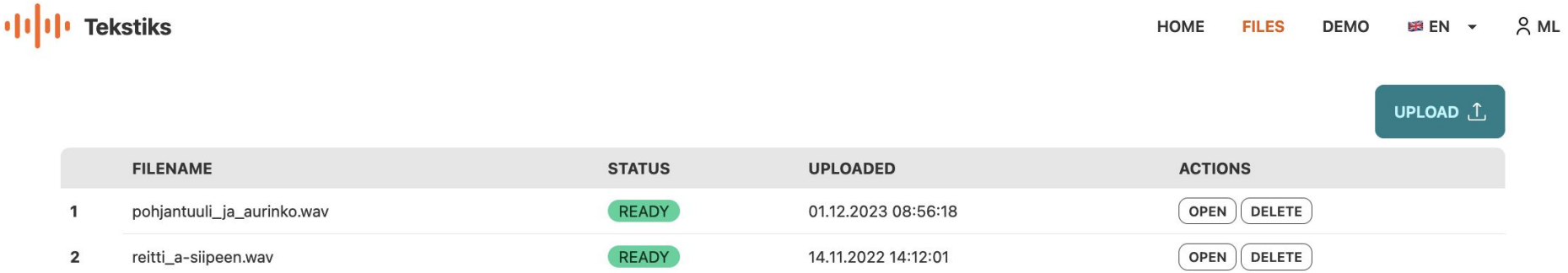
The editing of the transcription is interactive. The integrated audio player and the text are in sync. The currently playing word in coloured to simplify the manual editing.

4. Download the result

Download the transcription, currently the DOCX format is supported.

Tools:
tekstiks.ee


Tekstiks
HOME
FILES
DEMO
EN
ML
UPLOAD
FILENAME
STATUS
UPLOADED
ACTIONS
1
reitti_a-siipeen.wav
OPEN
DELETE
Upload a speech recording
Language
Finnish
File
Selaa...
pohjantuuli_ja_aurinko.wav
• Supported formats: wav, mp3, ogg, mp2, m4a, mp4, flac, amr, mpg.
• File size limit is 100MB.
UPLOAD
Tools:
tekstiks.ee



Tools:
tekstiks.ee

pohjantuuli_ja_aurinko.wav

Kesto: 00:00:36 Ladattu: 1. joulukuuta 23

Muokkaustila:
- (•) Tavallinen (korjaa tunnistusvirheet)
- ( ) Merkinnät (teksti ja puhe)

LATAA

S1 00:01
pohjantuuli ja aurinko pohjantuuli ja aurinko väittelivät kummalla olis enemmän voimaa

S1 00:06
kun he samalla näkivät kulkijan jolla oli yllään lämmin takki silloin he sopivat että se on voimakkaampi joka nopeammin saa kulkijan riisumaan takkinsa

S1 00:16
voihan tuuli alkoi puhaltaa niin että viuhuu mutta mitä kovempaa se puhalsi sitä tarkemmin kääri miestäkin ympärilleen

S1 00:23

1.0x

pohjantuuli_ja_aurinko.wav

Kesto: 00:00:36 Ladattu: 1. joulukuuta 23

Muokkaustila:
- (•) Tavallinen (korjaa tunnistusvirheet)
- ( ) Merkinnät (teksti ja puhe)

LATAA

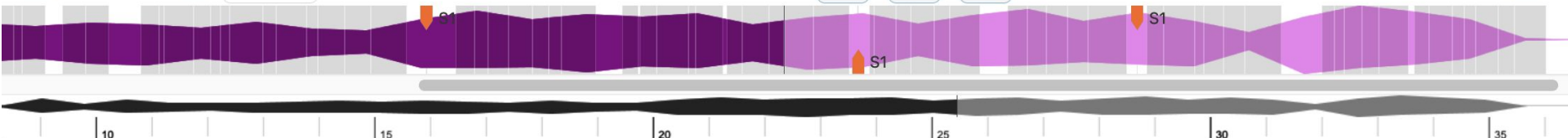
S1 00:01
pohjantuuli ja aurinko pohjantuuli ja aurinko väittelivät kummalla olis enemmän voimaa

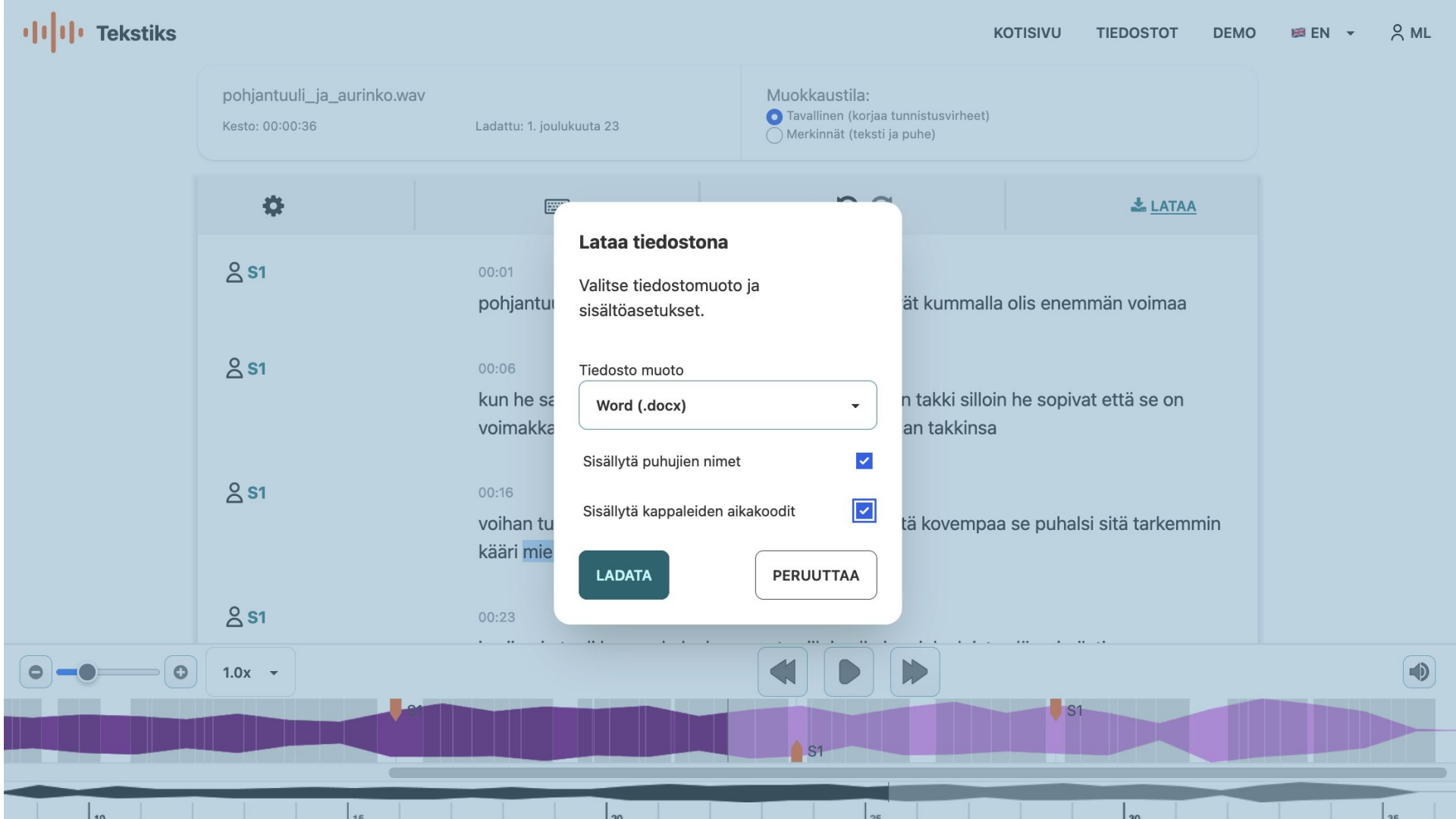
S1 00:06
kun he samalla näkivät kulkijan jolla oli yllään lämmin takki silloin he sopivat että se on voimakkaampi joka nopeammin saa kulkijan riisumaan takkinsa

S1 00:16
voihan tuuli alkoi puhaltaa niin että viuhuu mutta mitä kovempaa se puhalsi sitä tarkemmin kääri mies takin ympärilleen

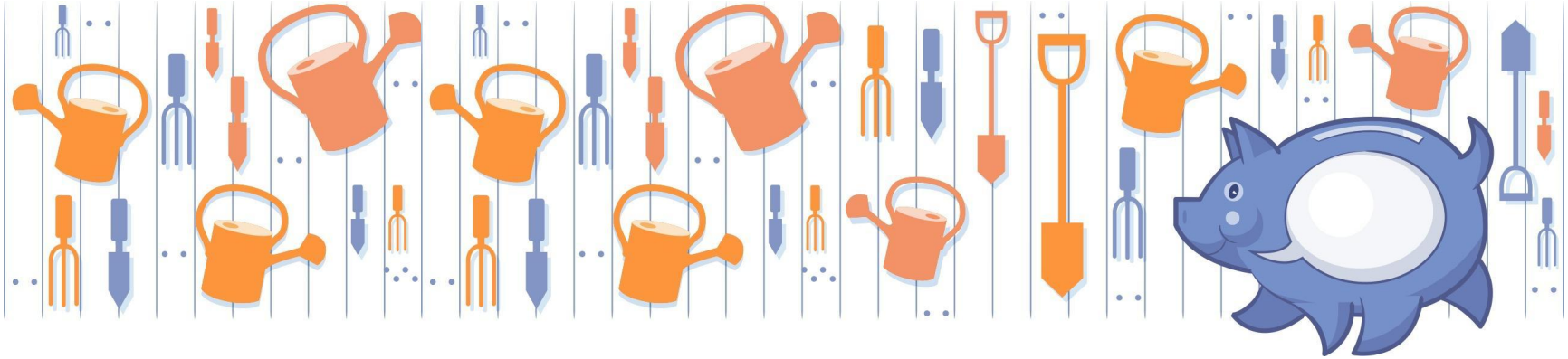
S1 00:23

1.0x




Tekstiks
KOTISIVU
TIEDOSTOT
DEMO
EN
ML
pohjantuuli_ja_aurinko.wav
Kesto: 00:00:36
Ladattu: 1. joulukuuta 23
Muokkaustila:
Tavallinen (korjaa tunnistusvirheet)
Merkinnät (teksti ja puhe)
LATAA
S1
00:01
S1
00:06
S1
00:16
S1
00:23
Lataa tiedostona
Valitse tiedostomuoto ja sisältöasetukset.
Tiedosto muoto
Word (.docx)
Sisällytä puhujien nimet
Sisällytä kappaleiden aikakoodit
LADATA
PERUUTTAA
1.0x

KIELIPANKKI
The Language Bank of Finland



WP2.1:

SOCIAL DATA SCIENCE

FIN-CLARIN

What kind of data can be deposited?

- Text or speech in any natural language
  - text documents, audio and video recordings
  - annotations and transcripts may be included
  - must be in useful and well-described formats
- Make sure you have the rights to distribute the data, at least for research purposes
  - Copyright and other Intellectual Property Rights
  - Personal data

Contact FIN-CLARIN for details

fin-clarin@helsinki.fi

Submit the basic details about your dataset

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?

The home organization of

Are safety measures required for sharing?

- The Language Bank of Finland can offer various means for protecting personal data and other types of restricted content:
  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 (first resource: findarc)
  4. Sensitive Data (SD) services at CSC

Two types of agreements are usually needed:

1. Deposition (distribution) license agreement (DELA)
   https://www.kielipankki.fi/support/dela/

   - The Language Bank (University of Helsinki) obtains the right to distribute the resource to end-users under specific terms and conditions.
   - a DELA is usually required, unless the University of Helsinki is the rightholder or the resource is already available under a public license.
   - For resources containing personal data, the agreement also includes the details of the data controller and the resource-specific data protection terms and conditions **regarding the redistribution of the data via the Language Bank**.

2. End-User License Agreement (EULA)
   - The End-User agrees to use the data under specific conditions which the rightholder has approved.

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

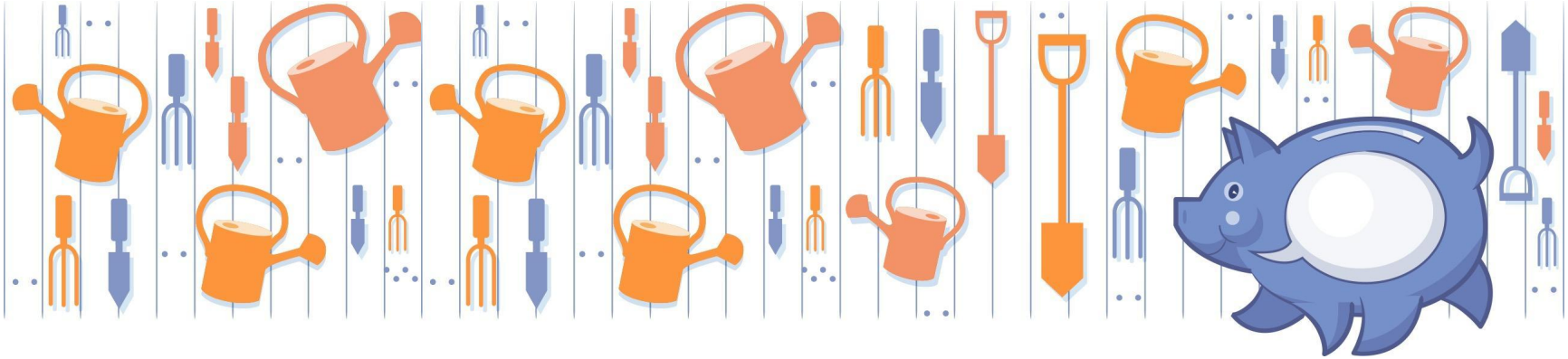
+BY the author(s) must be cited
+NC non-commercial use only
+ID login is required
+PLAN a research plan is required
+**PRIV** **contains personal data**
+NORED redistribution is not allowed
+DEP modified versions can be redistributed via CLARIN

and other resource-specific conditions, if required (e.g., data protection terms and conditions)

Are you using a resource that includes personal data (license condition **+PRIV**)?

- As a user, you are required to submit the title of your project and a link to the **public privacy notice** regarding your project, to be published by Kielipankki.
  - The privacy notice must be submitted when applying for access to a restricted resource that contains personal data.
  - If the resource is available without an application, submit your privacy notice via the e-form.

KIELIPANKKI
The Language Bank of Finland



WP2.2:

LEARNERS' ASSESSMENT ENVIRONMENTS

FIN-CLARIN

Donate Speech Corpus v1.0

http://urn.fi/urn:nbn:fi:lb-2020090321

Tweet #lb_puhelahjat

Donate Speech datasets (puhelahjat) for research use

Suomeksi

Donate Speech datasets for commercial use: see further details on another page

Important information for all users of this resource: Removal requests

Versions of this resource:	
Donate Speech Corpus, version 1.0 Metadata License (for researchers) Attribution instructions	Apply for access rights, academic research use only **+PRIV:** This resource contains personal data. Submit public information about personal data processing Download the resource
Donate Speech Corpus: Sample	Download the resource

LAHJOITA PUHETTA

Pulistaan meille parempia palveluja

Nyt kerätään kaikenlaista puhuttua suomea! Lahjoittamasi puheen puheen avulla esimerkiksi ääniohjatut laitteet voivat oppia ymmärtämään erilaisia murteita ja puhetapoja. Valitse alta aihe, josta haluat puhua.

ELÄINYSTÄVÄT

SUKELLA KESÄÄN

KERRO KORONA-AJASTA

URHEILUHETKET

LÄHELLÄNI JUURI NYT

Lahjoittaminen onnistuu parhaiten mobiilisovelluksella:

Lataa App Storesta

GET IT ON Google Play

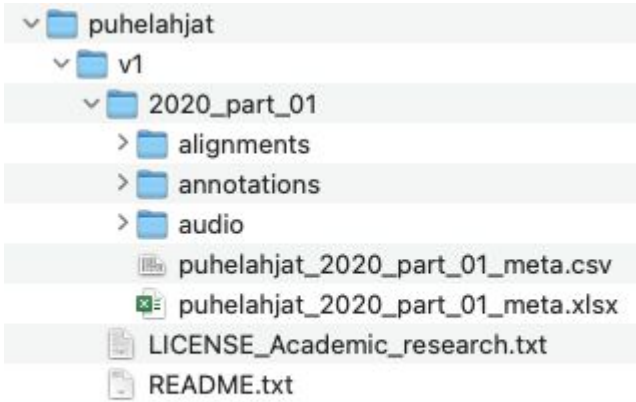
KIELIPANKKI
The Language Bank of Finland

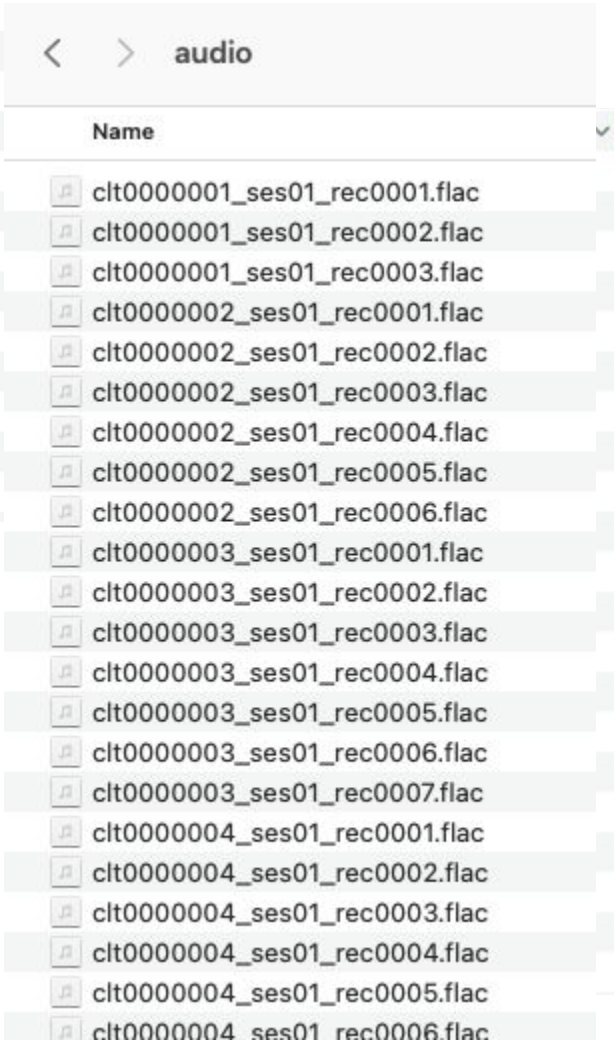
LATAUKSET
DOWNLOADS

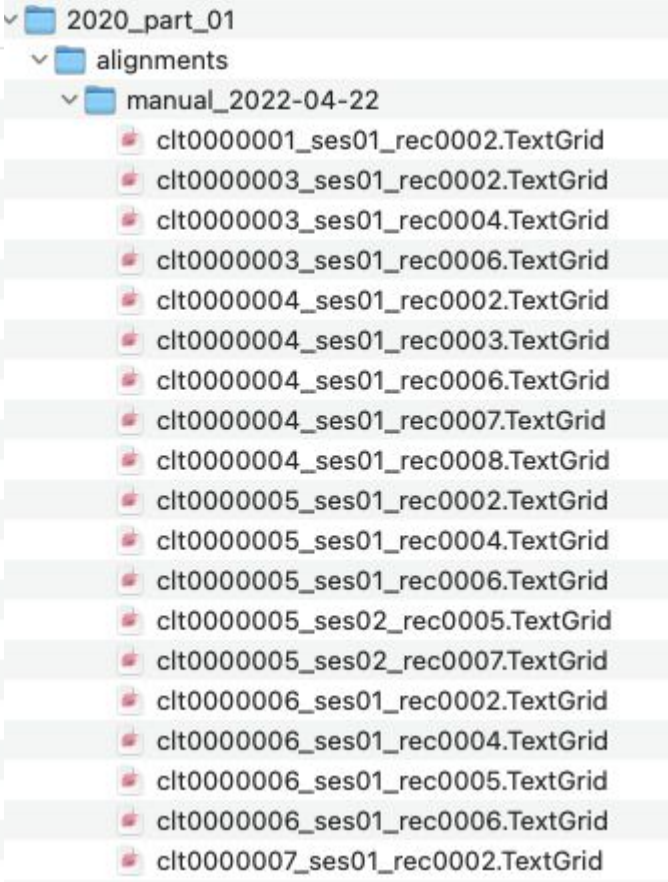
HOME UP

Location: /download/puhelahjat/complete/ Logged in as:

Name	Size	Description
puhelahjat_v1_2020_part_01_alignments_manual_2022-04-22.zip	11M	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_01_annotations_manual_2021-10-28.zip	3.3M	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_01_audio.zip	11G	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_01_meta.zip	252K	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_02_alignments_manual_2022-04-22.zip	11M	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_02_annotations_manual_2021-10-28.zip	3.1M	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_02_audio.zip	11G	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_02_meta.zip	239K	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_03_alignments_manual_2022-04-22.zip	10M	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_03_annotations_manual_2021-10-28.zip	3.1M	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_03_audio.zip	11G	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_03_meta.zip	268K	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_04_alignments_manual_2022-04-22.zip	11M	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_04_annotations_manual_2021-10-28.zip	3.2M	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_04_audio.zip	11G	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_04_meta.zip	253K	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_05_alignments_manual_2022-04-22.zip	10M	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_05_annotations_manual_2021-10-28.zip	3.1M	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_05_audio.zip	11G	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_05_meta.zip	250K	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_06_alignments_manual_2022-04-22.zip	11M	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_06_annotations_manual_2021-10-28.zip	3.1M	Donate Speech Corpus 1.0
puhelahjat_v1_2020_part_06_audio.zip	11G	Donate Speech Corpus 1.0


puhelahjat
v1
2020_part_01
alignments
annotations
audio
puhelahjat_2020_part_01_meta.csv
puhelahjat_2020_part_01_meta.xlsx
LICENSE_Academic_research.txt
README.txt


audio
Name
clt0000001_ses01_rec0001.flac
clt0000001_ses01_rec0002.flac
clt0000001_ses01_rec0003.flac
clt0000002_ses01_rec0001.flac
clt0000002_ses01_rec0002.flac
clt0000002_ses01_rec0003.flac
clt0000002_ses01_rec0004.flac
clt0000002_ses01_rec0005.flac
clt0000002_ses01_rec0006.flac
clt0000003_ses01_rec0001.flac
clt0000003_ses01_rec0002.flac
clt0000003_ses01_rec0003.flac
clt0000003_ses01_rec0004.flac
clt0000003_ses01_rec0005.flac
clt0000003_ses01_rec0006.flac
clt0000003_ses01_rec0007.flac
clt0000004_ses01_rec0001.flac
clt0000004_ses01_rec0002.flac
clt0000004_ses01_rec0003.flac
clt0000004_ses01_rec0004.flac
clt0000004_ses01_rec0005.flac
clt0000004_ses01_rec0006.flac


2020_part_01
alignments
manual_2022-04-22
clt0000001_ses01_rec0002.TextGrid
clt0000003_ses01_rec0002.TextGrid
clt0000003_ses01_rec0004.TextGrid
clt0000003_ses01_rec0006.TextGrid
clt0000004_ses01_rec0002.TextGrid
clt0000004_ses01_rec0003.TextGrid
clt0000004_ses01_rec0006.TextGrid
clt0000004_ses01_rec0007.TextGrid
clt0000004_ses01_rec0008.TextGrid
clt0000005_ses01_rec0002.TextGrid
clt0000005_ses01_rec0004.TextGrid
clt0000005_ses01_rec0006.TextGrid
clt0000005_ses02_rec0005.TextGrid
clt0000005_ses02_rec0007.TextGrid
clt0000006_ses01_rec0002.TextGrid
clt0000006_ses01_rec0004.TextGrid
clt0000006_ses01_rec0005.TextGrid
clt0000006_ses01_rec0006.TextGrid
clt0000007_ses01_rec0002.TextGrid

KIELIPANKKI
The Language Bank of Finland

The Donate Speech Corpus has helped in training automatic speech recognizers

- Automatic speech recognition systems, trained by Aalto Speech Research team on the Donate Speech Corpus, are available on Zenodo.
- More on the content of the Donate Speech Corpus:

Moisio, A., Porjazovski, D., Rouhe, A. et al. Lahjoita puhetta: a large-scale corpus of spoken Finnish with some benchmarks. *Lang Resources & Evaluation* (2022). https://doi.org/10.1007/s10579-022-09606-3

FIN-CLARIN

L2 speech corpora (soon available!)

- Speech recorded from second-language learners in the DigiTala project (UHEL, Aalto, JYU), including oral skills assessment data:
  - DigiTala: L2 Finnish data from upper secondary schools and university, autumn 2021
  - DigiTala: L2 Finnish data from upper secondary schools, spring 2021
  - DigiTala: L2 Swedish data from adult language learners, spring 2023
  - DigiTala's YKI data (*Yleiset kielitutkinnot*, general language tests)

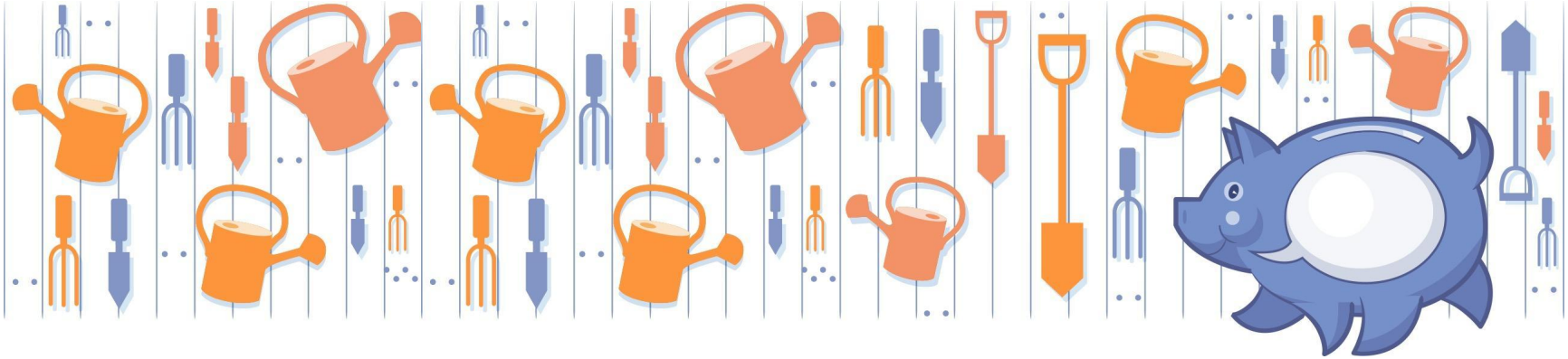
KIELIPANKKI
The Language Bank of Finland

ASR systems for L2 Finnish

- Models trained by Aalto Speech Research team for performing automatic speech recognition (ASR) and pronunciation rating for children's L2 Finnish.
  - The systems are available on HuggingFace Hub.
- Tools for the automatic speaking assessment of spontaneous L2 Finnish and Swedish (DigiTala project)

FIN-CLARIN

KIELIPANKKI
The Language Bank of Finland



WP2.3:

TRANSLATION AND INTERPRETATION

FIN-CLARIN


Finnish | Swedish | Other language | Parallel
Log in Suomi | Svenska | English 11 Cite Korp MENU
KORP v9
Christmas Gospel text-to-speech (kpv-... selected — 269 of 3.16G tokens
KIELIPANKKI The Language Bank of Finland
[word = "lap.*"]
Extended
Livvi-Karelian
word
begins with
laps
Aa
tai
ja
Valinnat
Add token
Add boundary
Add lang
Search
Parallel corpus with linked audio:
Christmas Gospel text-to-speech in four Uralic languages, Korp
KWIC: hits per page: 10
sort within corpora: not sorted
Statistics: compile based on: word
Show statistics
KWIC
Statistics
Results: 15
« ‹ 1 2 › »
Go to page of 2
Show context
CHRISTMAS GOSPEL TEXT-TO-SPEECH (KPV-KRL-MYV-OLO) (BETA)
i häi sai poijan, enzimäzen lapsen . Häi kabaloičči lapsen da pani žiivatoin soimeh, ku heile ei olluh sijua ma
Сійӧ чужтіс Пи , ассыс медводдза кагасӧ . Рузумӧ гартыштӧмӧн Мария пуктіс кагасӧ лясниӧ , ӧд туйм
i häi sai poijan, enzimäzen lapsen. Häi kabaloičči lapsen da pani žiivatoin soimeh, ku heile ei olluh sijua matkuniekoin talois.
Сійӧ чужтіс Пи , ассыс медводдза кагасӧ . Рузумӧ гартыштӧмӧн Мария пуктіс кагасӧ лясниӧ , ӧд туйморт керкаын налы
Tämä on teile sen merkinny: työ lövvättö lapsen , kudai kabaloitunnu viruu soimes. »
Лясниысь ті аддзанныд рузумӧ гартыштӧм кагаӧс – тайӧ лоӧ тіянлы пасӧн » .
Hyö kiirehel lähtiettih i löyttih Marija, Josif da lapsi , kudai virui soimes.
Найӧ тэрмасьӧмӧн мунісны сэтчӧ да аддзисны сэтысь Марияӧс , Ӧсипӧс да лясниын куйлысь кагаӧс .
Heijän nähtyy paimoit ruvettih sanelemah, midä heile oli sanottu lapseh näh.
Сэсся висьталісны кага йылысь кывлӧмторсӧ .
CORPUS
Christmas Gospel text-to-speech (olo) (beta)
Metadata
Cite corpus
Persistent identifier:
urn:nbn:fi:lb-2023111601
Licence: CC-BY-NC
TEXT ATTRIBUTES
ISO 639 code: olo
verse: :LUK.2.7:
WORD ATTRIBUTES
base form: lapsi
baseform (compound boundaries):

Add lang | Search

KWIC: hits per page: 10 | sort within corpora: not sorted | Statistics: compile based on: word | Show statistics

KWIC | Statistics

Results: 15

« ‹ 1 2 › » Go to page of 2 Show context

CHRISTMAS GOSPEL TEXT-TO-SPEECH (KPV-KRL-MYV-OLO) (BETA)

i häi sai poijan, enzimäzen **lapsen** . Häi kabaloičči lapsen da pani žiivatoin soimeh, ku heile ei olluh sijua mat

Сійӧ чужтіс Пи , ассыс медводдза кагасӧ . Рузумӧ гартыштӧмӧн Мария пуктіс кагасӧ лясниӧ , ӧд туймо

i häi sai poijan, enzimäzen lapsen. Häi kabaloičči **lapsen** da pani žiivatoin soimeh, ku heile ei olluh sijua matkuniekoin talois.

Сійӧ чужтіс Пи , ассыс медводдза кагасӧ . Рузумӧ гартыштӧмӧн Мария пуктіс кагасӧ лясниӧ , ӧд туйморт керкаын налы уз

Tämä on teile sen merkinny: työ lövvättö **lapsen** , kudai kabaloitunnu viruu soimes. »

Лясниысь ті аддзанныд рузумӧ гартыштӧм кагаӧс – тайӧ лоӧ тіянлы кагасӧн » .

Hyö kiirehel lähtiettih i löyttih Marija, Josif da **lapsi** , kudai virui soimes.

Найӧ тэрмасьӧмӧн мунісны сэтчӧ да аддзисны сэтысь Мариянӧс , Ӧсипӧс да лясӧн куйлысь кагаӧс .

Heijän nähtyy paimoit ruvettih sanelemah, midä heile oli sanottu **lapseh** näh.

Сэсся висьталісны кага йылысь кывлӧмторсӧ .

CHRISTMAS GOSPEL TEXT-TO-SPEECH (KRL) (BETA)

i häi sai poijan, enzimäzen **lapsen** . Häi kabaloičči lapsen da pani žiivatoin soimeh, ku heile ei olluh sijua mat

Hiän šai Pojan , enšimmäisen Lapšen . Hiän kapaloičči Lapšen ta pani Hänet šoimeh , šentäh kun heilä ei löyty

i häi sai poijan, enzimäzen lapsen. Häi kabaloičči **lapsen** da pani žiivatoin soimeh, ku heile ei olluh sijua matkuniekoin talois.

Hiän šai Pojan , enšimmäisen Lapšen . Hiän kapaloičči Lapšen ta pani Hänet šoimeh , šentäh kun heilä ei löytyn tilua mat

Tämä on teile sen merkinny: työ lövvättö **lapsen** , kudai kabaloitunnu viruu soimes. »

Täštä työ tunnetta Hänet : ripakkoloih kiäritty Lapši makuau šoimešša .

Hyö kiirehel lähtiettih i löyttih Marija, Josif da **lapsi** , kudai virui soimes.

Hyö lähettih kiirehešti Viflejemih ta löyvettih šielä Muarie , Joosefi ta Lapši , kumpani makasi šoimešša .

Heijän nähtyy paimoit ruvettih sanelemah, midä heile oli sanottu **lapseh** näh.

Tämän nähtyö paimenet kerrottih , mitä anhelit šanottih heilä Lapšešta .

« ‹ 1 2 › » Go to page of 2

Download hit page as…

CORPUS

Christmas Gospel text-to-speech (olo) (beta)

Metadata
Cite corpus
Persistent identifier: urn:nbn:fi:lb-2023111601
Licence: CC-BY-NC

TEXT ATTRIBUTES

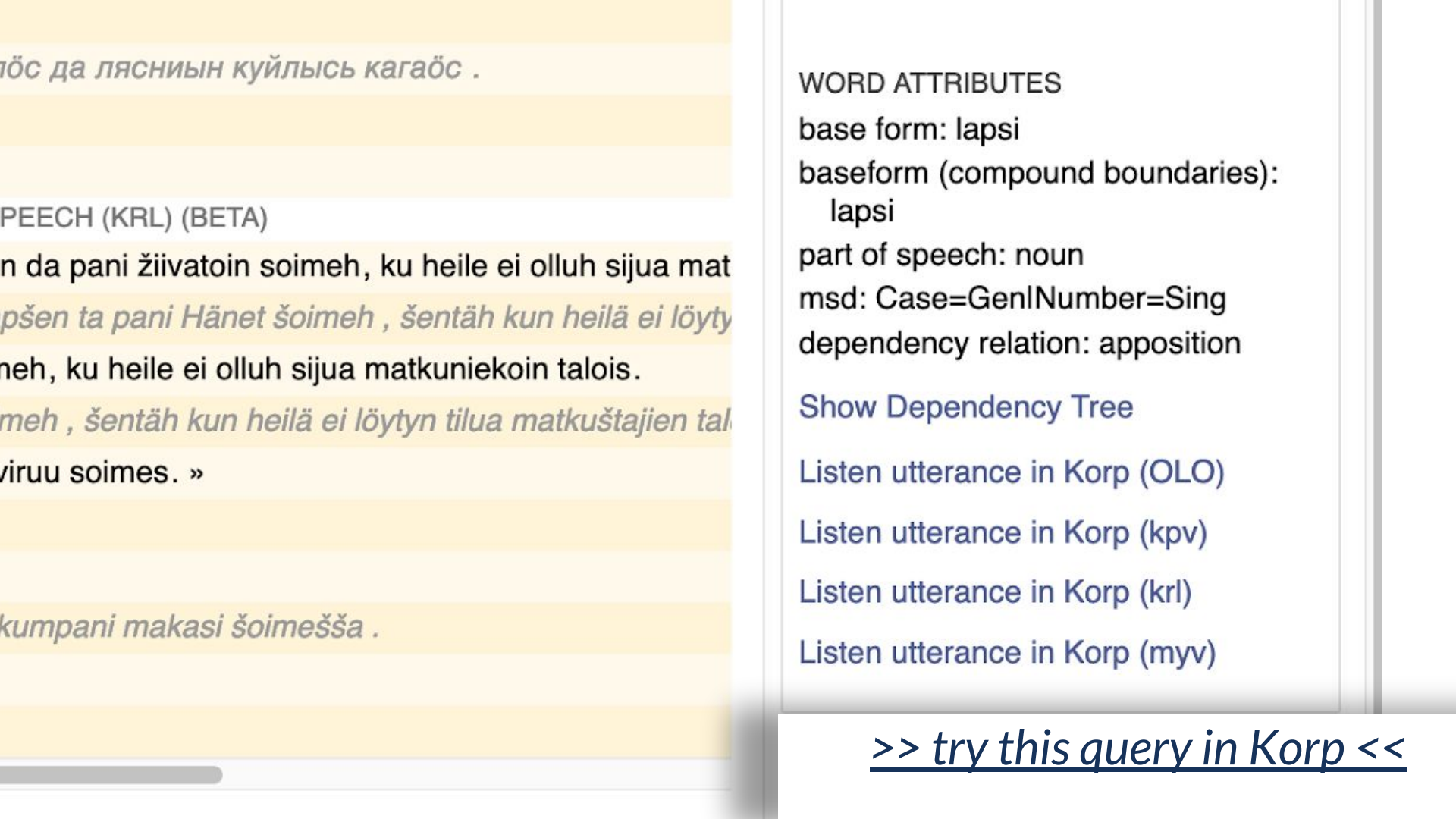
ISO 639 code: olo
verse: :LUK.2.7:

WORD ATTRIBUTES

base form: lapsi
baseform (compound boundaries): lapsi
part of speech: noun
msd: Case=Gen|Number=Sing
dependency relation: apposition

Show Dependency Tree

Listen utterance in Korp (OLO)
Listen utterance in Korp (kpv)
Listen utterance in Korp (krl)
Listen utterance in Korp (myv)


лöс да лясниын куйлысь кагаöс .
PEECH (KRL) (BETA)
n da pani žiivatoin soimeh, ku heile ei olluh sijua mat
pšen ta pani Hänet šoimeh , šentäh kun heilä ei löyty
meh, ku heile ei olluh sijua matkuniekoin talois.
meh , šentäh kun heilä ei löytyn tilua matkuštajien tal
viruu soimes. »
kumpani makasi šoimešša .
WORD ATTRIBUTES
base form: lapsi
baseform (compound boundaries): lapsi
part of speech: noun
msd: Case=Gen|Number=Sing
dependency relation: apposition
Show Dependency Tree
Listen utterance in Korp (OLO)
Listen utterance in Korp (kpv)
Listen utterance in Korp (krl)
Listen utterance in Korp (myv)
>> try this query in Korp <<

KIELIPANKKI
The Language Bank of Finland

LATAUKSET
DOWNLOADS

HOME UP

Location: /download/xmas-gospel-tts/ Logged in as:

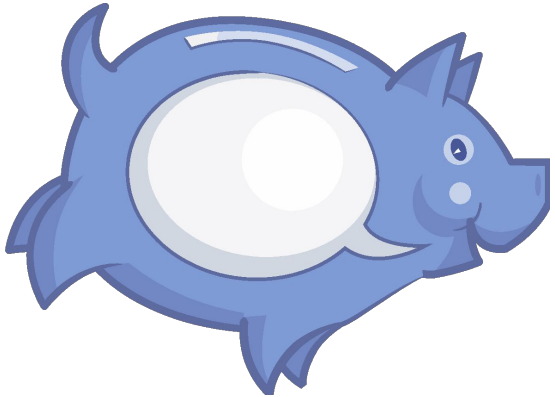
Name	Size	Description
LICENSE.txt	19K	Licence information
README.txt	2.0K	More information
xmas-gospel-tts-src.zip	38M	CC BY NC

CLARIN PUB- tai ACA-lisensoidut valikoidut versiot ladattavista kieliain
hakemistosta /appl/data/kielipankki/. Palvelimien käyttö edellyttää CSC-

Selected versions of the downloadable CLARIN PUB or ACA licensed co
/appl/data/kielipankki/. To use the servers you need a CSC account. Tech

Parallel corpus with linked audio:
Christmas Gospel text-to-speech in four Uralic languages, source
(= downloadable version)

KIELIPANKKI
The Language Bank of Finland



Kiitos! Tack! Thank you!

www.kielipankki.fi

General support
fin-clarin@helsinki.fi

Technical support
kielipankki@csc.fi

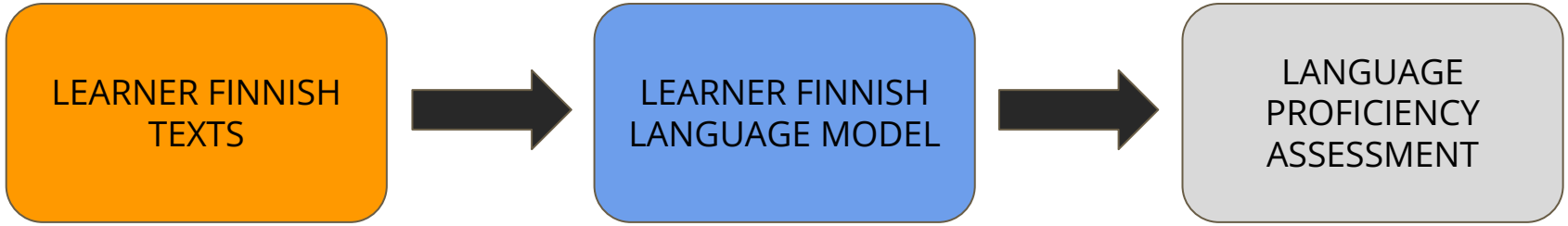
WP 2.4
The Helsinki Term Bank for the Arts and Sciences

FIN-CLARIAH Meeting in Tampere, December 1st, 2023

Harri Kettunen (harri.kettunen@helsinki.fi) &
Tiina Onikki-Rantajääskö (tiina.onikki@helsinki.fi)
tieteentermipankki-info@helsinki.fi

TIETEEN TERMIPANKKI
VETENSKAPSTERMBANKEN I FINLAND
THE HELSINKI TERM BANK FOR THE ARTS AND SCIENCES

WP2.5 Automated Text Tools for Learner Language



- trained with 4 CEFR-annotated data sets (CEFR = Common European Framework of Reference, from A1 to C2).
  - ICLFI: International Learner Corpus of Finnish
  - LAS2: The Advanced Finnish Learners' Corpus
  - CEFLING: Linguistic Basis of the Common European Framework for L2 English and L2 Finnish
  - TOPLING – Paths in Second Language Acquisition
    - = 7700 texts, 1.6 million tokens
- trained on FinBERT (by Turku NLP group)

Contact:
Jenny Tarvainen: jenny.h.tarvainen@jyu.fi
Ida Toivanen: ida.m.toivanen@jyu.fi
Ari Huhta: ari.huhta@jyu.fi

PRELIMINARY PROGRAM

11.00-11.10 Welcoming Words by Sanna Kumpulainen, Associate Professor in Information Studies, Tampere University

11.10-12.00 Keynote I on Studying SSH Research Needs: Elina Late, Senior Research Fellow in Information Studies, Tampere University

12.00-13.00 Lunch

13.00-13.45 Keynote II on Language Models: Sampo Pyysalo, Associate Professor at the Department of Computing, University of Turku

13.45-14.30 Work Package Presentations I

13.45-14.00 WP1.3 ~~Veronika Laippala~~: Noise-Tolerant NLP

14.00-14.20 WP 1.1, 1.2, 2.1, 2.2, 2.3 Mietta Lennes: Kielipankki – The Language Bank of Finland

14.20-14.25 WP2.4 Harri Kettunen: Helsinki Term Bank for the Arts and Sciences

14.25-14.30 WP2.5 Jenny Tarvainen: Automated Text Tools for Learner Language

14.30-15.00 Coffee

15.00-16.00 Work Package Presentations II

15.00-15.05 WP3.1 Martin Matthiesen: Pipeline from the National Library to CSC

15.05-15.10 WP3.2 Tanja Välisalo: Named Entity Recognition for NARC Data

15.10-15.15 WP4.3 Eetu Mäkelä: Evaluation and Subsetting

15.15-15.20 W4.1 Julia Matveeva: Metadata Harmonization

15.20-15.25 WP4.4 Mikko Laitinen: Twitter

15.25-15.30 WP4.2 Eero Hyvönen: LOD

15.30-15.35 WP3.4 Raine Koskimaa: Game Streams

15.35-15.40 WP3.3 Maria Valaste: Qualitative Surveys

15.40-15.45 WP3.5 Kimmo Elo (Risto Turunen replacing): Text Networks

15.45-15.50 WP5 Sanna Kumpulainen: Evidence-based RI Development + Education & Resources

16.00-17.00 Free Chilling & Refreshments / Parallel session: Executive Board Meeting (with Zoom option)

WP 3.1

Data of the National Library

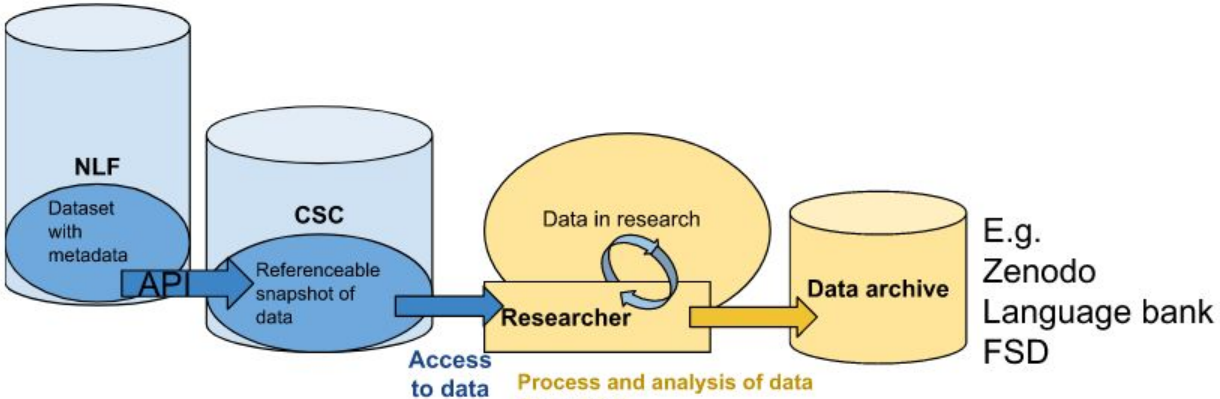
- Contains of out-of-copyright digitized publications
- Newspapers (1771-1918), journals (1771-1912), books
    - news, articles, advertisements, discussion, weather, … etc
- Page text in ALTO XML, Access images (JPG)
- 9 TB (zip compressed)

Added value for researchers

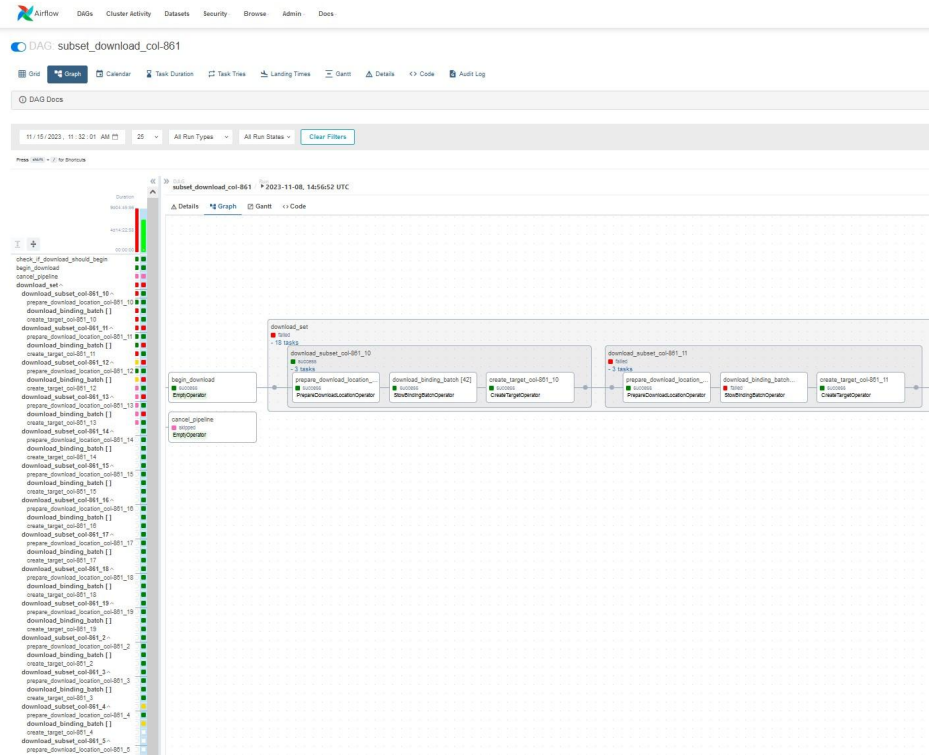
- **A**ccessibility: Dataset is closer to CSC processing environments (Puhti, Mahti, LUMI). We handle:
  - catalogue discovery
  - metadata parsing
  - API access
  - downloading tens of millions of files.
- **I**nteroperability: Different research groups can be sure they use the same data
- **R**eproducibility: Datasets can be referenced, older versions can be stored.
- See http://urn.fi/urn:nbn:fi:lb-202311261

Dataflow

Metadata description
Digitization
Formats



Download control

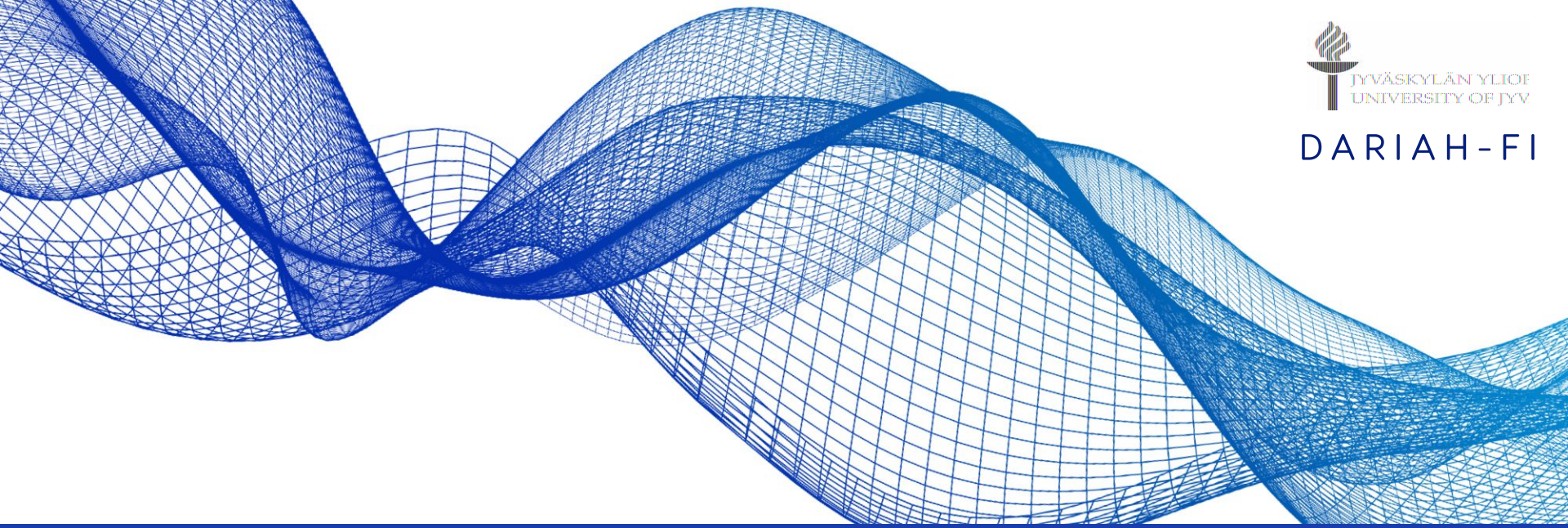


Lessons learned

- Do not underestimate scaling
  - “Harmless” access errors
  - Airflow memory requirements
- Well defined APIs have value
- Less is more
  - squashfs in containers vs zip files

Future work / Open questions

- Defining the prerequisites for offering copyrighted material in the future
- Making the reference URN machine actionable
  - E.g.: http://urn.fi/urn:nbn:fi:lb-202311261#a0c1b2
- Long-term availability
  - Ongoing research needed to keep data available
  - Lost snapshots cannot be recovered, only the last version of the data
  - Download of full dataset takes several weeks



JYVÄSKYLÄN YLIOP
UNIVERSITY OF JYV

DARIAH-FI

WP3.2 Named entity recognition (NER) for state authority archival data

Antero Holmila, Venla Poso, Ida Toivanen, Tanja Välisalo
In co-operation with the National Archives of Finland / DALAI project

The usability of mass digitised state authority data in the National Archives of Finland was improved by developing an AI based model for Named Entity Recognition

- **First prototype** of the model published by NARC and can be tested at arkkiivi.fi and https://huggingface.co/Kansallisarkisto/finbert-ner
- **Final model to be published in Spring 2024**

Current unpublished NER model

- obtains comparable results with non-OCR’d data, while
- **significantly improving** results with OCR’d state authority archival data

FUTURE

Automated **per-unit** metadata identification with NARC data:

- **Document date:** utilizing current work with NER of journal number (diaarinumero) detection
- **Document content type** (*e.g. correspondence, report*) using language tools and computer vision

WP4.3 EVALUATION AND SUBSETTING

WP4.3 EVALUATION AND SUBSETTING - ACHIEVEMENTS

Developed a prototype for “universal” data analysis and subsetting pipeline:

- Researched and adopted a sustainable and customizable software package (Elasticsearch & Kibana stack)
- The software is usable on multiple levels of expertise: API and UI, ‘clickable’ data browsing interface and more serious programmatic digital humanities application.
- Developed practices for easily integrating new datasets into the solution for relatively rapidly adding new datasets for future projects.
- Developed (and still developing) documentation for easy adaptation and duplication of the process: Deploying the software on CSC or similar hosting service, adding and extracting datasets.
- Tested the solution pipeline in a practical project: https://receptionreader.com/latin

WP4.3 EVALUATION AND SUBSETTING - CHALLENGES

Challenges (and lessons learned)

- Developing complex plugins from scratch quite a demanding process - this work is still in progress.
- Even “ready to use” harmonized data usually needs some further fine refinement before being ready for integration into a database.
- Adopting into use in any hosting service always seems to need some amount of jumping through hoops. But as a positive surprise: here, relatively little.

- Positive lessons learned: The system is way more usable almost out of the box than we expected: Serves as a very smooth and responsive backend for the test use case (Latin reuse reader).
- The standard data analysis tools are excellent for getting a quick overview of a dataset.

WP4.3 EVALUATION AND SUBSETTING - FUTURE

Future use & steps for this solution:

- Adopting Elasticsearch & Kibana in place of our current custom data API - for use by various research groups in Helsinki.
- Developing language analyzer plugins to better serve scientific use.
- Smoothing out the workflow leading from data indexing to analysis and extraction towards a standardized model based on our test use cases.

WP4.1 METADATA HARMONIZATION

WP4.4 TWITTER (X)

- Five massive datasets collected
  - NTS: 799 million words (c. 74m messages from 888,098 accounts)
    - A user interface for easy data access + data shared through Kielipankki
  - Network datasets
    - 19,345 ego networks with 759,495 nodes
    - C. 11 billion words
    - Nordic region, UK, US and Australia (broad geographic coverage)

- The biggest challenge was obviously closing down the API > unique datasets from 2006-May2023;
- [We cannot solely rely on eccentric billionaires and the Big Tech for data access]
- Digital Single Markets directive and the subsequent changes in the Finnish copyright law
- Next up: data enrichment for background variables (age, social class in addition to networks)

FIN-CLARIAH Meeting, Dec 1, 2023, Tampere

WP 4.2 National Linked Open Data Infrastructure: Sampo Data Services & Semantic Portals

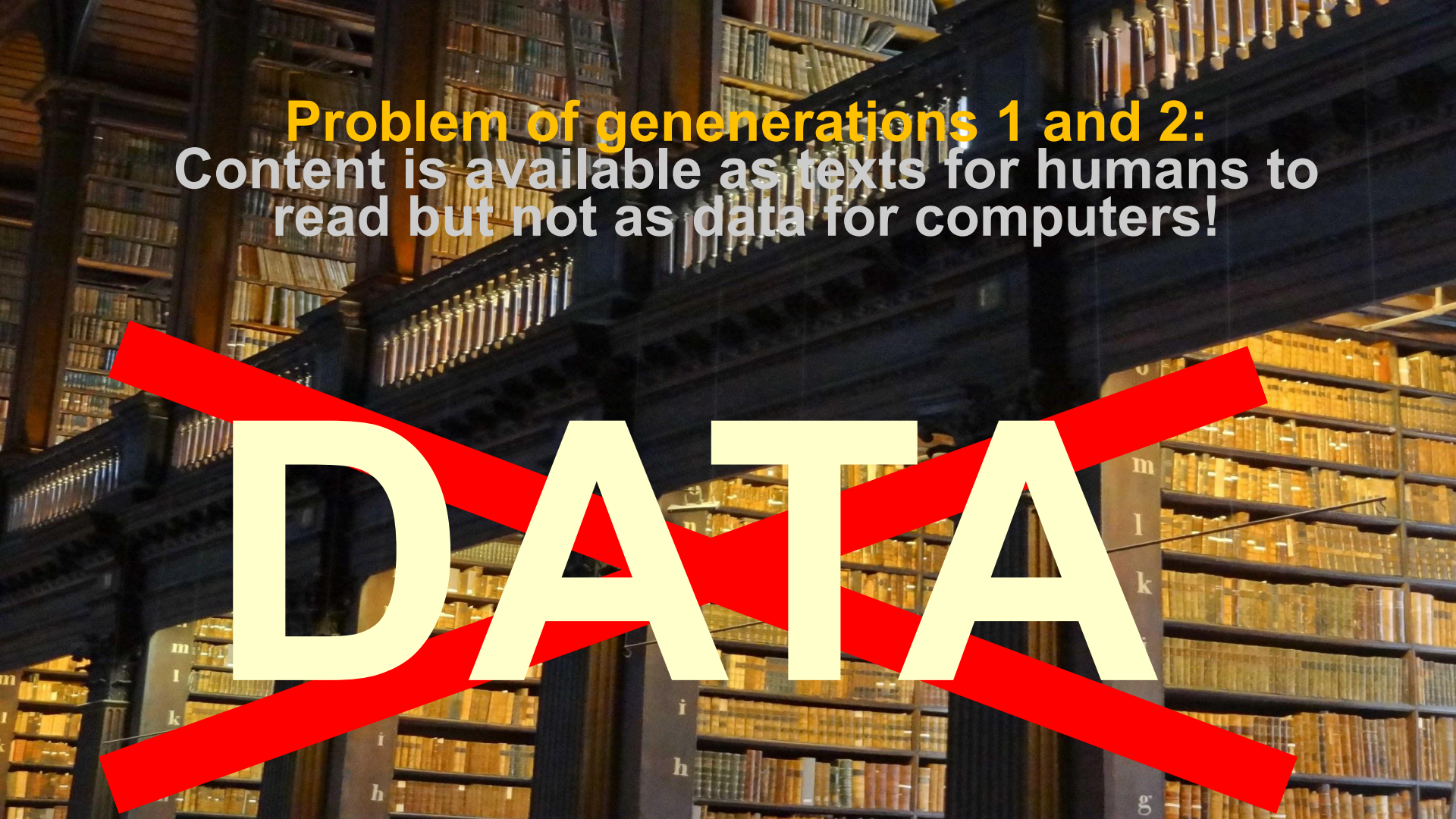
Eero Hyvönen
Professor, Department of Computer Science, Aalto University
Director, Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki
https://seco.cs.aalto.fi/u/eahyvone/

We Aim at a Paradigm Change:
4 Generations of Publishing Data for Humanities

1. Texts (Engravings, Hand-written, and Printed)
2. Online Systems for Searching and Exploring
3. Publishing Content as Linked Data with Tools for DH
4. Automatic Knowledge Discovery and Artificial Intelligence

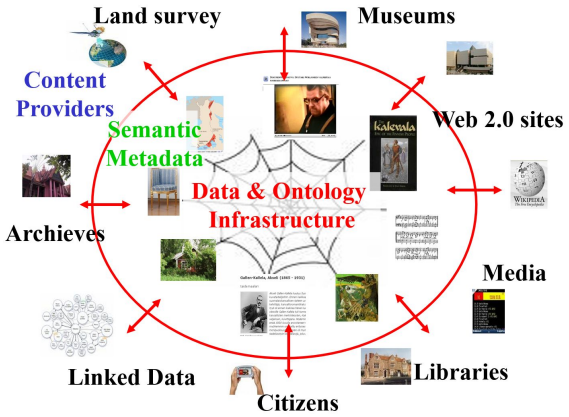
Our Focus

E. Hyvönen, : Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery, *Semantic Web*, 1 (11), 2020 pdf

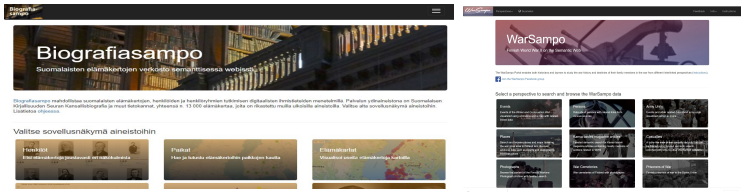

Problem of genenerations 1 and 2:
Content is available as texts for humans to read but not as data for computers!
DATA

Using Sampo Model for Cultural Heritage Data by 1) human users and 2) machines

Sampo LOD Service



Sampo Portals & Applications



Digital Humanities Research Tools

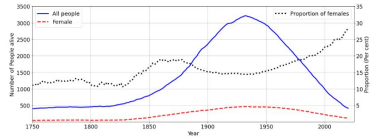


Fig. 3: Amount of people alive on timeline

Fig. 4: Average lifespan of biographee's

Linked Data Finland Service:
https://ldf.fi

Elements of a National Semantic Web Infrastructure for Open Cultural Heritage FAIR Data

Applications

Sampo-systems, Finna, Europeana, ARIADNEplus, TimeMachine…

Ontologies
Ontology services
Finto.fi

Datasets
Data services
LDF.fi

Human Infrastructure

Linked Open Data Infrastructure for Finnish CH Cloud

Standards
Best practices

W3C, …

Shared data models

DC, CIDOC-CRM, …

Software tools
Sampo-UI, …

Eero Hyvönen: How to Create a National Cross-domain Ontology and Linked Data Infrastructure and Use It on the Semantic Web. *Semantic Web*, forth-coming, 2023.

Sampo Series of LOD Services & Portals

A. Gallen-Kallela: The Defense of the Sampo, 1896, Turku Art Museum,

Sampo-UI

1. **MuseumFinland** – Finnish Museums on the Semantic Web (2004) **[39 000 users]**
2. **CultureSampo** – Finnish Culture on the Semantic Web (2008) **[107 000 users]**
3. **TravelSampo** - Mobile Contextualized Services of Cultural Tourism (2011)
4. **BookSampo** – Finnish Fiction Literature on the Semantic Web (2011) **[1.6 million users in 2022]**
5. **WW1LOD** – World War I Linked Open Data (2014)
6. **WarSampo –** Finnish World War 2 on the Semantic Web (2015-19) **[1 100 000 users]**
7. **Norssi Alumni on the Semantic Web** – Historical person registry using LOD (2017)
8. **U.S. Congress Prosopographer –** U.S. Congress Legislators 1789-2018
9. **BiographySampo** - Finnish Biographies on the Semantic Web (2018-20) **[381 000 users]**
10. ~~**NameSampo** – Linked Data Workbench for Toponomastic Research (2019) **[55 000 users]**~~
11. **WarVictimSampo 1914-1922** – National War History **[80 000 users]**
12. **Mapping Manuscript Migrations (MMM)** – medieval manuscripts (2020) **[9100 users]**
13. **AcademySampo –** Finnish Academic People 1640 – 1899 (2021) **[26 000 users]**
14. **FindSampo** – Archaeological Finds on the Semantic Web (2021) **[7 000 users]**
15. **WarMemoirSampo** – Memoirs of Finnish WW2 veterans (2021) **[3 800 users]**
16. **LetterSampo** – Early Modern Letters on the Semantic Web (2022)
17. **ParliamentSampo** – Parliament of Finland on the Semantic Web (2023)
18. **LawSampo –** Finnish Legislation and Case Law on the Semantic Web (2023)
19. **BookSampo II** – Semantic Search, Browsing, and Data-analyses (2023)
20. **OperaSampo** – Opera and music theater performances in Finland 1830-1960 (2023)
21. **ConfermentSampo** – Conferment Ceremonies of the University of Helsinki 1640-1899 (2023)

IN USE

Year of 5 Sampos 2023

Sampo = Mythical artifact of the Finnish Epic Kalevala that gives to its owner riches and good fortune. ***A metaphore of amazing ancient technology.***
Links to all Sampos: https://seco.cs.aalto.fi/applications/sampo/

Eero Hyvönen: **Digital Humanities on the Semantic Web: Sampo Model and Portal Series**. *Semantic Web*, vol. 14, no. 4, pp. 729-744,, 2023. pdf link

Project homepage: https://seco.cs.aalto.fi/projects/semparl/

PARLAMENTTISAMPO AVAA EDUSKUNNAN MILJOONA PUHETTA JA KANSANEDUSTAJIEN VERKOSTOT KAIKKIEN TUTKITTAVIKSI

Eduskunnan täysistunnoissa on vuosina 1907–2022 pidetty lähes miljoona puheenvuoroa, ja yhteensä puhujia on ollut noin 2 800. Uusi Parlamenttisampo-palvelu tarjoaa kaiken tämän aineiston yhtenäisenä datana, mikä mahdollistaa kansanedustajien sekä poliittisen kielen, kulttuurin ja verkostojen tutkimisen entistä helpommalla tavalla.

Lähes miljoona eduskunnan täysistunnoissa pidettyä puheenvuoroa on Parlamenttisampo-hankkeessa ensimmäistä kertaa muunnettu linkitetyksi avoimeksi dataksi (englanniksi *Linked Open Data*). Avoimen datapalvelun päälle on kehitetty semanttinen Parlamenttisampo-portaali, jota voivat tutkijoiden lisäksi käyttää media, poliitikot ja suuri yleisö.

Parlamenttisampo on uusi jäsen Sampo-portaalien sarjassa, joilla on ollut jopa miljoonia käyttäjiä semanttisessa webissä. Muita Sampo-portaaleja ovat esimerkiksi lainsäädäntöä ja oikeustapauksia kokoava *Lakisampo.fi* ja *Biografiasampo.fi*, joka sisältää Suomalaisen Kirjallisuuden Seuran Kansallisbiografiassa julkaistujen suomalaisten merkkihenkilöiden elämäkerrat ja heidän verkostoja.

EDUSKUNNAN AINEISTOT AVOIMEKSI FAIR-DATAKSI JA SOVELLUKSIKSI

Suomalaisen demokratian yksi perusta on, että eduskunnan päätöksenteko ja lainsäädäntötyö ovat avoimia ja niihin liittyvät aineistot saatavilla. Keskeinen tähän liittyvä

Pitkäaikainen kansanedustaja Veikko Vennamo (1913–1997) käytti eduskunnan täysistunnoissa enemmän puheenvuoroja kuin kukaan muu. Vuonna 1974 hänet jopa kannettiin ulos istuntosalista, kun hän ei suostunut poistumaan täysistunnosta puhemiehen kehotuksesta huolimatta.

Tieteessä tapahtuu, 41 (1), 16.2.2023
https://seco.cs.aalto.fi/publications/2023/hyvonen-parlamenttisampo-tt-2023.pdf

And More Sampos are Being Forged at the Smithy

A. Gallen-Kallela: The Forging of the Sampo, 1893, Ateneum Museum,

- **CoCoSampo** – Finnish letters on the Semantic Web 1809–1917
  - https://seco.cs.aalto.fi/projects/coco/
- **FinEst-LawSampo** – Cross-border Multilingual Legislation Search
  - https://seco.cs.aalto.fi/projects/finestsampo/
- **PASampo** – British Museum Portable Antiquity Scheme case study
  - https://seco.cs.aalto.fi/projects/diginuma/
- **CoinSampo** – Coin finds of the Nartional Museum of Finland
  - https://seco.cs.aalto.fi/projects/diginuma/
- **NomismaSampo** – Numismatics based on Nomisma.org model and data
  - https://seco.cs.aalto.fi/projects/diginuma/
- **ArtSampo** – Finnish Art on the Semantic Web
  - https://seco.cs.aalto.fi/projects/taidesampo/

- **SampoSampo** – Finnish Linked Open Data Cloud of Cultural Heritage

https://seco.cs.aalto.fi/projects/fin-clariah/

WP3.4 GAME STREAMS

-Twitch Chat Collector Tool Ready

-both live streams and recorded videos

-Analysis Tool (Discussion Dynamics, Uniqueness Classification)

-Functional, some polishing needed

-Publication of code in Github (under which organization?)

-Publication of collected data, permission from Twitch required

https://drive.google.com/file/d/1hS4t4XSE6KQNvmDFOx8RQHqN-Xd629bN/view?usp=sharing

WP3.3 QUALITATIVE SURVEYS

WP members: Adeline Clarke (University of Helsinki), Ida Toivanen (University of Jyväskylä), Jani-Matti Tirkkonen (University of Eastern Finland), Jaakko Peltonen (Tampere University) and Maria Valaste (University of Helsinki)

- Github:
  - R package: https://github.com/DARIAH-FI-Survey-Concept-Network/finnsurveytext
    - Website: https://dariah-fi-survey-concept-network.github.io/finnsurveytext/
  - Other materials: https://github.com/DARIAH-FI-Survey-Concept-Network/
- Our video:
  - finnsurveytext video
  - If you want to comment, please contact us! (adeline.clarke@helsinki.fi)

Funded by
the European Union
NextGenerationEU

UNIVERSITY OF TURKU

WP 3.5: Text network analysis of political texts

WP presentation 01.12.2023

Adjunct professor Kimmo Elo

Senior Research Fellow, Centre for Parliamentary Studies

E-mail: kimmo.elo@utu.fi

Funded by
the European Union
NextGenerationEU

UNIVERSITY OF TURKU

Summary

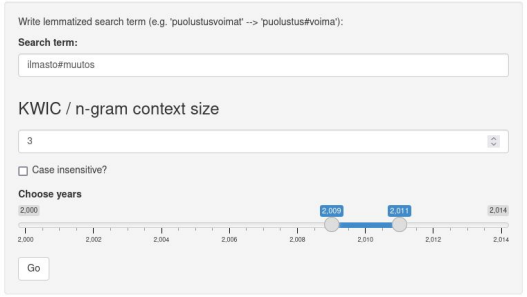
- **Core data:** FinParl (plenary debates of the Finnish *eduskunta* since 1905)
- **Deliverables:**
  - **KWIC tool for FinParl corpus:** This tool provides a user interface to query word embeddings with KWIC method. The tool offers a simple, yet intuitive user interface built with R Shiny. **Output:** KWIC table, n-grams table, collocation network
  - **TNA tool for the analysis of speeches of Finnish MPs:** This tool provides functionalities for vocabulary based content analysis of political speeches. The user selects an MP and can then study 1) a timeline of the MP's plenary speeches, 2) a wordcloud of max. 500 most used words by the MP, as well 3) a speaker-to-concept network consisting of the 50 most frequently used concepts of the selected MP and and of his/her most similar colleagues by vocabulary
- **Release:** release beta versions will be opened for public use/testing on December 19, 2023

Funded by
the European Union
NextGenerationEU

UNIVERSITY OF TURKU

KWIC-tool (1/2)

Text networks



KWIC | n-gram

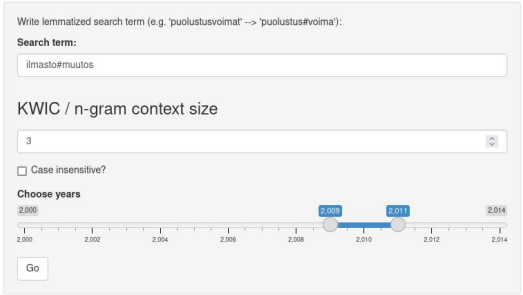
docname	from	to	pre	keyword	post	pattern
Fl2009_9_97.1	428	428	, finanssikriisi ja	ilmastonmuutoksen	aiheuttama lähtö entisiltä	ilmastonmuutoksen
Fl2009_6_4.2	215	215	. Mainitsen esimerkkeinä	ilmastonmuutoksen	, energian saatavuuden	ilmastonmuutoksen
Fl2009_6_7.1	79	79	Kansalliset rajat ylittävää	ilmastonmuutosta	on mahdotonta estää	ilmastonmuutosta
Fl2009_6_7.1	184	184	. Globaalit haasteet	ilmastonmuutoksesta	talouskriisiin ja edelleen	ilmastonmuutoksesta
Fl2009_6_8.1	217	217	syvenevä talouslama ja	ilmastonmuutos	. Näitä uhkia	ilmastonmuutos
Fl2009_6_9.2	44	44	. Globaalisti katsottuna	ilmastonmuutos	, ympäristötuhot ,	ilmastonmuutos
Fl2009_6_9.2	72	72	huomioitu , että	ilmastonmuutos	ja ympäristökysymykset voivat	ilmastonmuutos
Fl2009_6_13.2	119	119	, luonto ,	ilmastonmuutos	, influenssaepidemiat ,	ilmastonmuutos
Fl2009_33_7.1	414	414	mikä auttaa myös	ilmastonmuutoksen	vastaisessa taistelussa .	ilmastonmuutoksen
Fl2009_33_7.1	447	447	on erittäin perusteltu	ilmastonmuutoksen	näkökulmasta . Aivan	ilmastonmuutoksen
Fl2009_5_5.1	1382	1382	ja ilmastopolitiikkaa ja	ilmastonmuutokseen	sopeutumista . Maailmantalouden	ilmastonmuutokseen
Fl2009_5_10.3	103	103	tällä hetkellä ,	ilmastonmuutosta	. Voidaksemme jatkossa	ilmastonmuutosta
Fl2009_6_81.1	42	42	, ympäristöhaasteet ,	ilmastonmuutos	, luonnonkatastrofit —	ilmastonmuutos
Fl2009_6_97.1	111	111	bruttokansantuotteesta . Myös	ilmastonmuutoksen	torjunta edistää rauhantyötä	ilmastonmuutoksen
Fl2009_6_97.1	134	134	. Lisäksi kiihtyvä	ilmastonmuutos	aiheuttaa ankaraa kuivuutta	ilmastonmuutos
Fl2009_6_122.1	475	475	tarttuvista taudeista ,	ilmastonmuutoksesta	, suuronnettomuuksista ja	ilmastonmuutoksesta
Fl2009_6_123.1	97	97	mainittiin tällaisina uhkina	ilmastonmuutos	, energian ja	ilmastonmuutos
Fl2009_6_175.1	136	136	tuleekin huomioida niin	ilmastonmuutos	, luonnonkatastrofit ,	ilmastonmuutos
Fl2009_6_192.1	377	377	. Ympäristö ja	ilmastonmuutos	on mainittu tässä	ilmastonmuutos
Fl2009_6_192.1	394	394	. Ympäristöhaasteet ja	ilmastonmuutos	ovat omana kappaleena	ilmastonmuutos
Fl2009_7_138.1	328	328	kehitettävä . Myös	ilmastonmuutoksen	hillitseminen lisää julkisen	ilmastonmuutoksen
Fl2009_7_148.1	124	124	on huomioitava myös	ilmastonmuutoksen	torjunnan työmarkkinoille tuoma	ilmastonmuutoksen
Fl2009_7_152.1	160	160	näkökulmasta kestäviä ja	ilmastonmuutoksen	hillinnän kannalta tehokkaita	ilmastonmuutoksen
Fl2009_7_167.1	172	172	myös verotus torjuu	ilmastonmuutosta	, suojelee luonnonvaroja	ilmastonmuutosta
Fl2009_14_82.1	450	450	uudistaa verotusta jo	ilmastonmuutoksen	vuoksi . Tässä	ilmastonmuutoksen
Fl2009_17_112.1	237	237	aivan nykyisin puhuttavasta	ilmastonmuutoksesta	huolimatta niin kauan	ilmastonmuutoksesta
Fl2009_17_114.1	77	77	jo viittasi ,	ilmastonmuutoksen	kysymykseen , Sahelin	ilmastonmuutoksen
Fl2009_17_119.1	361	361	joita muun muassa	ilmastonmuutos	aiheuttaa jo tänä	ilmastonmuutos
Fl2009_17_122.1	53	53	tukemaan kehittyviä maita	ilmastonmuutoksen	vastaisessa kamppailussa .	ilmastonmuutoksen
Fl2009_17_122.1	64	64	puheenjohtaja Salolainen otti	ilmastonmuutoksen	vastaiset toimet esille	ilmastonmuutoksen
Fl2009_19_212.1	85	85	ja ennen muuta	ilmastonmuutoksen	kannalta kestävällä ja	ilmastonmuutoksen
Fl2009_21_39.1	87	87	riippumatta meistä —	ilmastonmuutoksesta	, joka tällä	ilmastonmuutoksesta
Fl2009_22_153.1	109	109	kestävällä tavalla ,	ilmastonmuutosta	hillitsevällä tavalla ,	ilmastonmuutosta
Fl2009_22_153.1	151	151	. Se on	ilmastonmuutoksen	ja ympäristöteknologian kehittäminen	ilmastonmuutoksen
Fl2009_26_65.1	463	463	joita teemme taistelussa	ilmastonmuutosta	vastaan , ja	ilmastonmuutosta

Funded by
the European Union
NextGenerationEU

UNIVERSITY OF TURKU

KWIC-tool (2/2)

Text networks



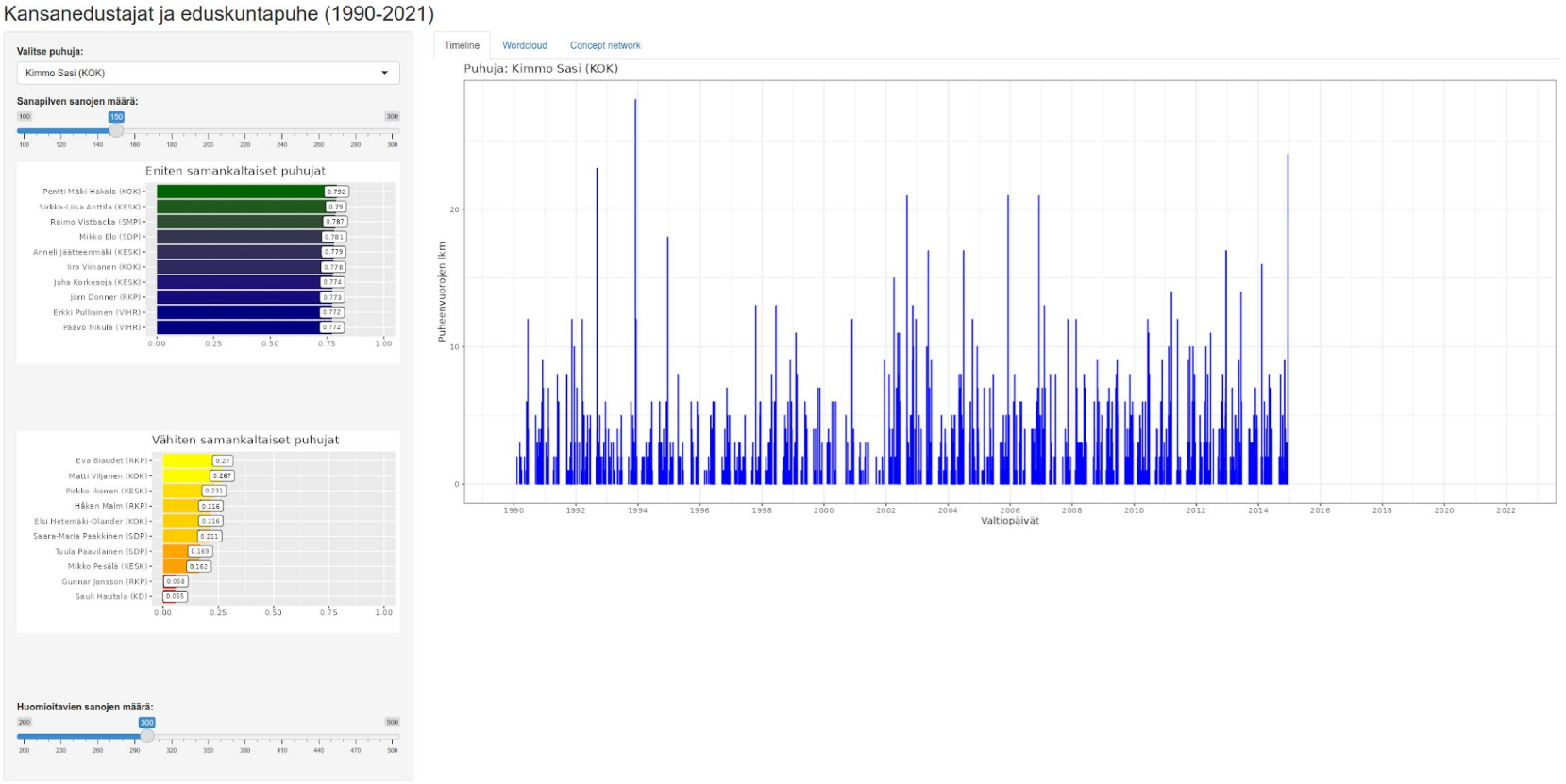
KWIC | n-gram

x	Freq
Arvoisa herra puhemies	181
on se että	158
ja se on	85
että se on	65
että meillä on	62
ja sitä kautta	54
Arvoisa rouva puhemies	51
on tärkeää että	44
Arvoisa puhemies On	43
se ei ole	36
sitä mieltä että	35
On tärkeää että	33
että Suomi on	31
myös se että	31
on hyvä että	30
On selvää että	30
on tällä hetkellä	30
tarkoittaa sitä että	30
Arvoisa puhemies Ed	28
ei ole varaa	28
Arvoisa puhemies Hallituksen	27
meillä ei ole	27
prosenttiin vuoteen mennessä	27
sillä tavalla että	27
Arvoisa puhemies Suomen	26
ja ennen kaikkea	26
ja tämä on	26
aikavälin ilmasto- ja	25
ei ole ollut	25
Se ei ole	25
niin että se	24
se että me	24
ei ole mitään	23
pitkän aikavälin ilmasto-	23
Suomi on sitoutunut	23

Funded by
the European Union
NextGenerationEU

UNIVERSITY OF TURKU

TNA-tool (1/3)

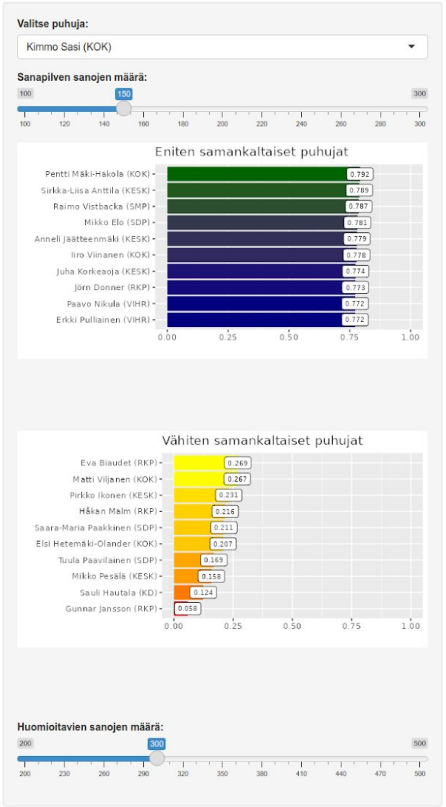


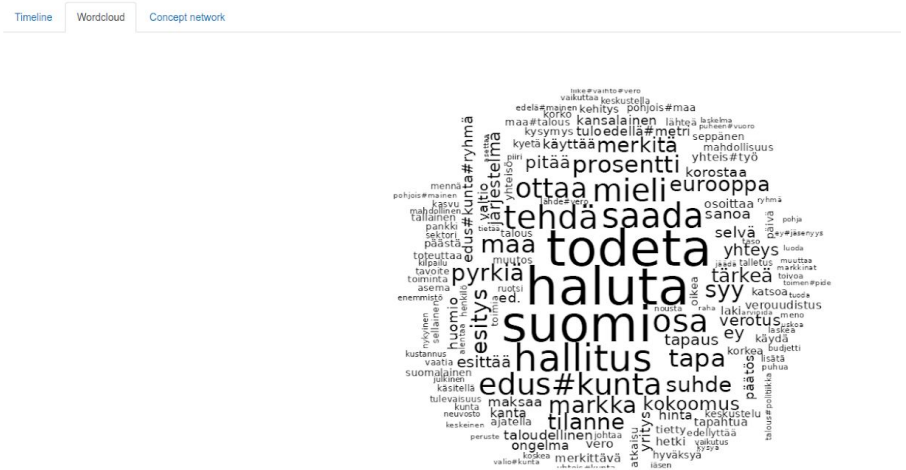
Funded by
the European Union
NextGenerationEU

UNIVERSITY OF TURKU

TNA-tool (2/3)

Kansanedustajat ja eduskuntapuhe (1990-2021)

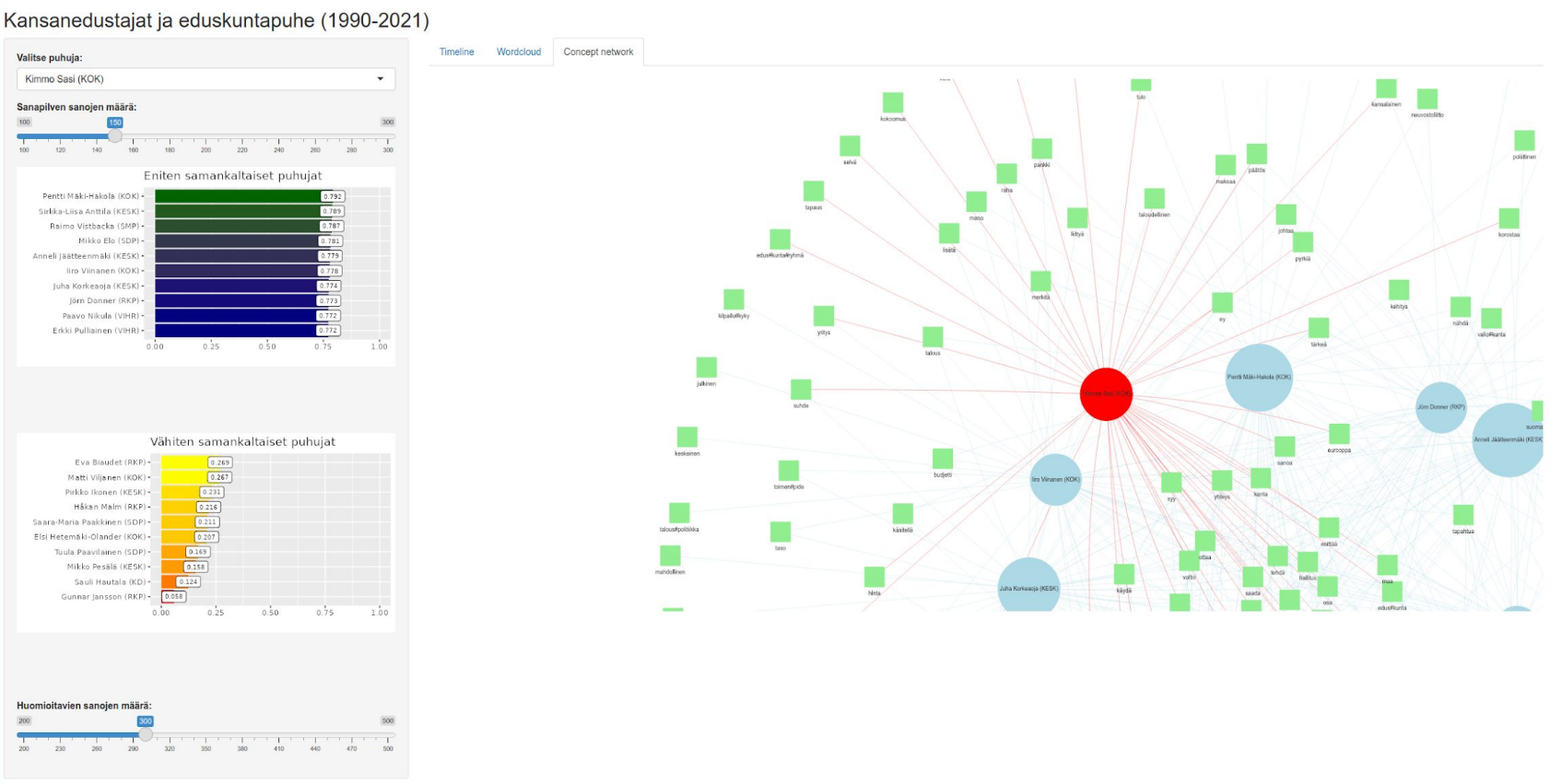




Funded by
the European Union
NextGenerationEU

UNIVERSITY OF TURKU

TNA-tool (3/3)



Tampere University

WP 5.1 and 5.2

Sanna Kumpulainen
Anna Sendra Toset
Farid Alijani
Jaakko Peltonen

Deliverables

- **D5.1.3: Protocol for collecting workshop data**
- **How to ensure that information flows from the users to the developers of the infrastructure?**
- https://doi.org/10.5281/zenodo.10217404

- **D5.2.2: Educational material**
- **Collection of resources related to the tools and materials shared in the DARIAH-FI**
- Dynamic document – published on DARIAH-FI website

- **D5.1.2: Log Data Collection and Analysis**
- Proof of concept of user-based recommendation
- https://github.com/mrgransky/DARIAH-FI

Publications

- Sendra Toset, A., Kumpulainen, S., & Late, E. (2023). *Putting the User in the Loop: Developing a Research Infrastructure for Social Sciences and Humanities Research*. 1119-1121. Poster session presented at 86th Annual Meeting of the Association for Information Science and Technology , London, United Kingdom. https://doi.org/10.1002/pra2.964
- Sendra Toset, A., Late, E. & Kumpulainen, S. (2023) More than data repositories: Perceived information needs for the development of social sciences and humanities research infrastructures. Information research. [Accepted for publication]

Future

- Tiny steps towards – an iterative process
- Mature infrastructure, i.e., organization and support services