New perspectives on archiving practices in language documentation

Niko Partanen, Michael Rießler & Marina Fedina

Lattice

cnrs

ENS

UNIVERSITÉ SORBONNE NOUVELLE PARIS 3

Who am I?

- I work currently as a visiting researcher in LATTICE laboratory, Paris
- I'm a linguist working with Uralic languages (minority languages in Russia)
- I'm one of the PI's in the project Language Documentation meets Language Technology: The Next Step in the Description of Komi
- I've been working with language documentation materials for last five years
- Other important collaborators are Michael Rießler (Freiburg), Rogier Blokland (Uppsala) and Marina Fedina (Syktyvkar)
- We try to adhere to field's best practices, but also foster more informal collaboration between different research projects

Talk structure

- Few words on language documentation and general context
- Research projects
- Our data
- Archive interaction
- “Ideal workflow”

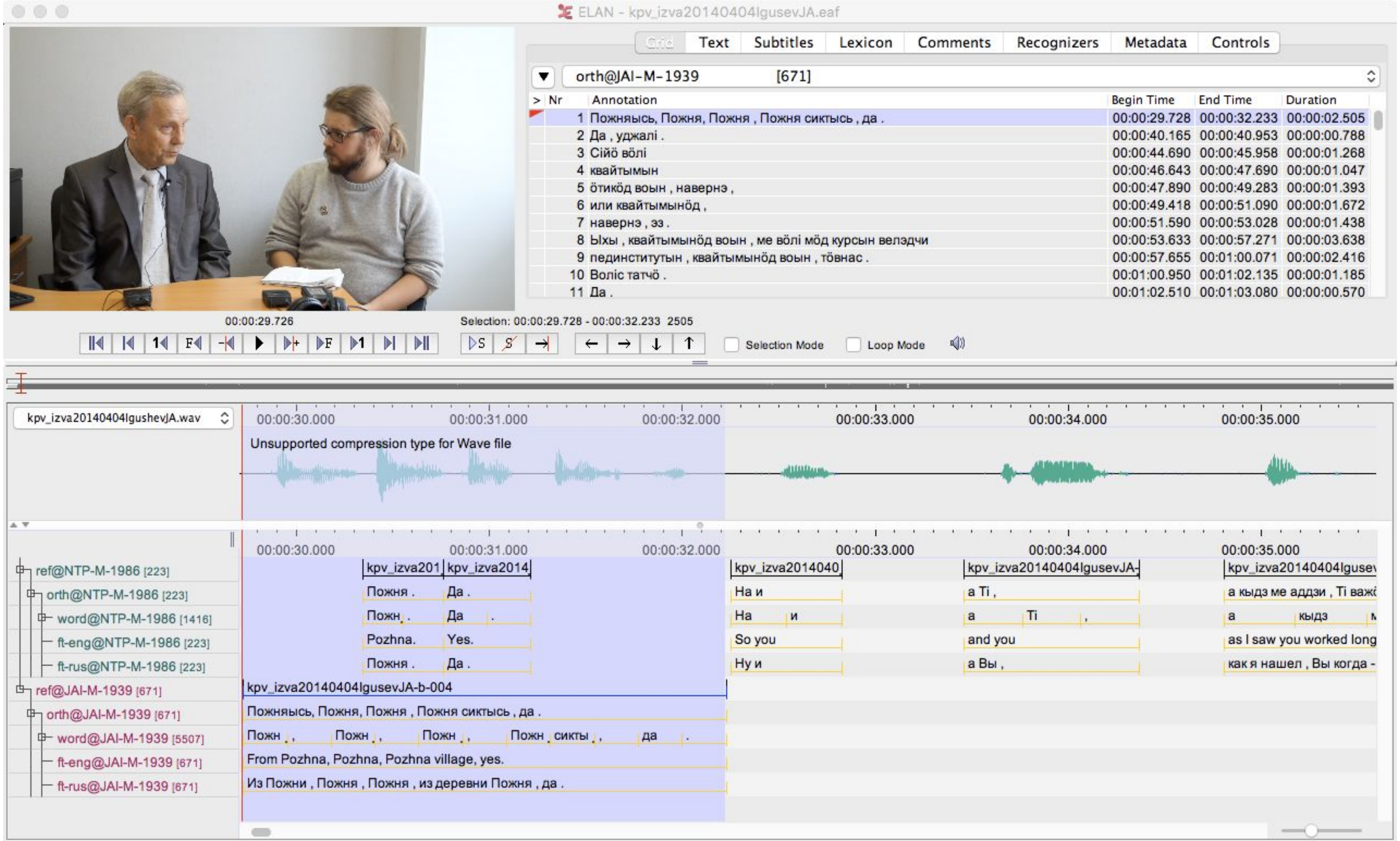
Slides available in: nikopartanen.github.io/IASA2017

Language documentation

- Language documentation ... concerned with the methods, tools, and theoretical underpinnings for compiling a representative and lasting **multipurpose** record of a language or one of its varieties (Himmelmann 1998; Himmelmann 2012)
- Usually focus on endangered languages
- Produces large amounts of linguistically annotated multimedia

- Traditionally lots of focus in metadata
  - Essential for linguistic research as well
  - The cultural context of recorded activities

So What
SO WHAT SO WHAT


ELAN - kpv_izva20140404IgusevJA.eaf
Grid
Text
Subtitles
Lexicon
Comments
Recognizers
Metadata
Controls
00:00:29.726
Selection Mode
Loop Mode
kpv_izva20140404IgushevJA.wav
Unsupported compression type for Wave file
00:00:30.000
00:00:31.000
00:00:32.000
00:00:33.000
00:00:34.000
00:00:35.000

orth@JAI-M-1939 [671]

Nr	Annotation	Begin Time	End Time	Duration
1	Пожнявысь, Пожня, Пожня , Пожня сиктысь , да .	00:00:29.728	00:00:32.233	00:00:02.505
2	Да , уджалі .	00:00:40.165	00:00:40.953	00:00:00.788
3	Сійӧ вӧлі	00:00:44.690	00:00:45.958	00:00:01.268
4	квайтымын	00:00:46.643	00:00:47.690	00:00:01.047
5	ӧтикӧд воын , наверно ,	00:00:47.890	00:00:49.283	00:00:01.393
6	или квайтымынӧд ,	00:00:49.418	00:00:51.090	00:00:01.672
7	наверно , ээ .	00:00:51.590	00:00:53.028	00:00:01.438
8	Ыхы , квайтымынӧд воын , ме вӧлі мӧд курсын велӧдчи	00:00:53.633	00:00:57.271	00:00:03.638
9	пединститутын , квайтымынӧд воын , тӧвнас .	00:00:57.655	00:01:00.071	00:00:02.416
10	Воліс татчӧ .	00:01:00.950	00:01:02.135	00:00:01.185
11	Да .	00:01:02.510	00:01:03.080	00:00:00.570

Selection: 00:00:29.728 - 00:00:32.233 2505

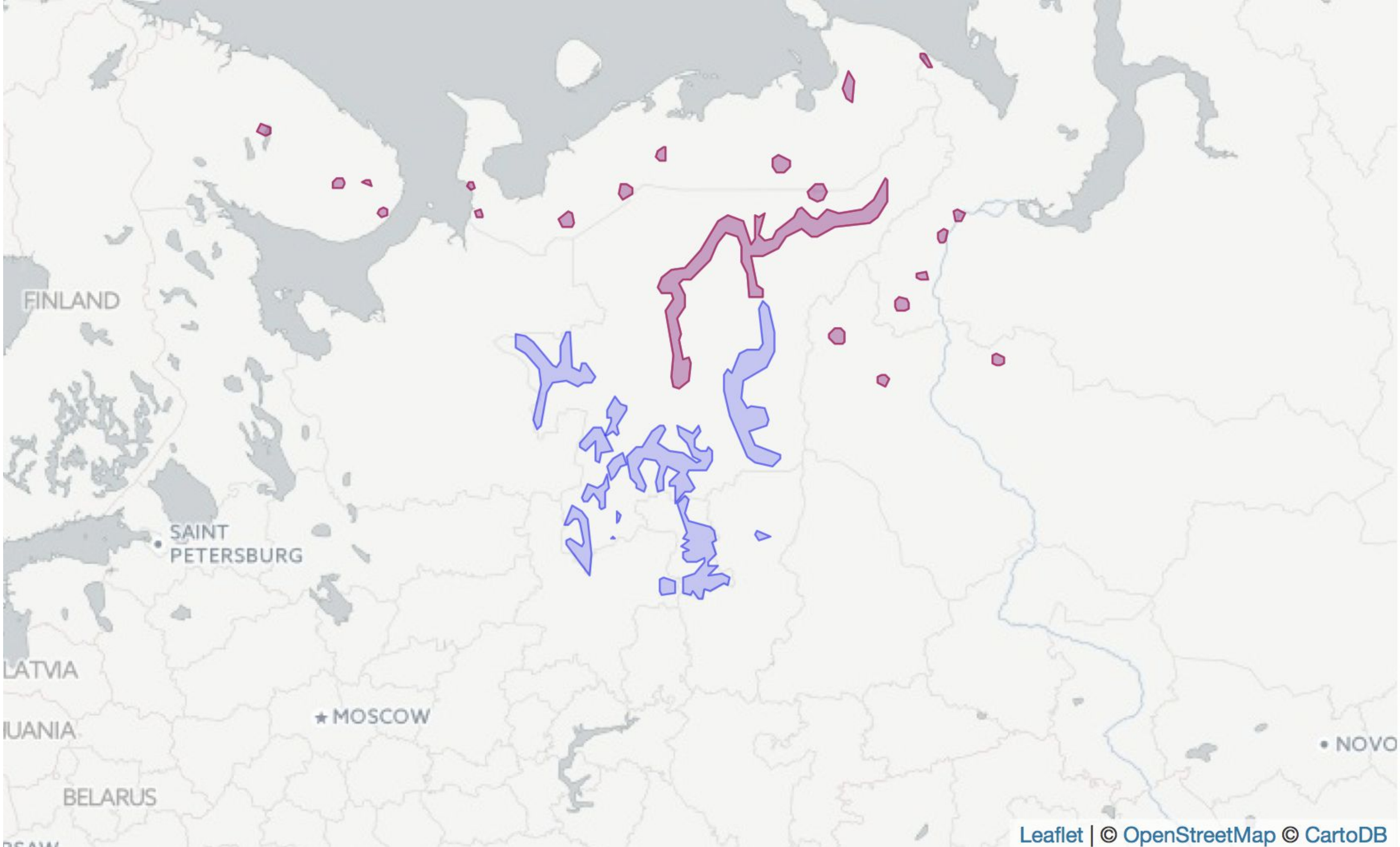
ref@NTP-M-1986 [223]		kpv_izva201	kpv_izva2014	kpv_izva2014040	kpv_izva20140404IgusevJA-
orth@NTP-M-1986 [223]		Пожня .	Да .	На и	а Ті ,
word@NTP-M-1986 [1416]		Пожн .	Да .	На и	а Ті ,
ft-eng@NTP-M-1986 [223]		Pozhna.	Yes.	So you	and you
ft-rus@NTP-M-1986 [223]		Пожня .	Да .	Ну и	а Вы ,
ref@JAI-M-1939 [671]	kpv_izva20140404IgusevJA-b-004				
orth@JAI-M-1939 [671]	Пожнявысь, Пожня, Пожня , Пожня сиктысь , да .				
word@JAI-M-1939 [5507]	Пожн , Пожн , Пожн , Пожн сикты , да .				
ft-eng@JAI-M-1939 [671]	From Pozhna, Pozhna, Pozhna village, yes.				
ft-rus@JAI-M-1939 [671]	Из Пожни , Пожня , Пожня , из деревни Пожня , да .				

kpv_izva20140404Igusev
а кыдз ме аддзи , Ті важ
а кыдз
as I saw you worked long
как я нашел , Вы когда -

Iźva Komi Documentation Project (1 & 2)

- Funded by Kone Foundation
- Collaborators in Uppsala, Freiburg, Syktyvkar and Paris
- First project in 2014-2016, the current in 2017-2019
- Data archived (mostly) into TLA in Nijmegen

- Initial focus in data collection and building the corpus
- Now we are working with language technology, corpus maintenance and improving the accessibility of the data, also conducting research on the data

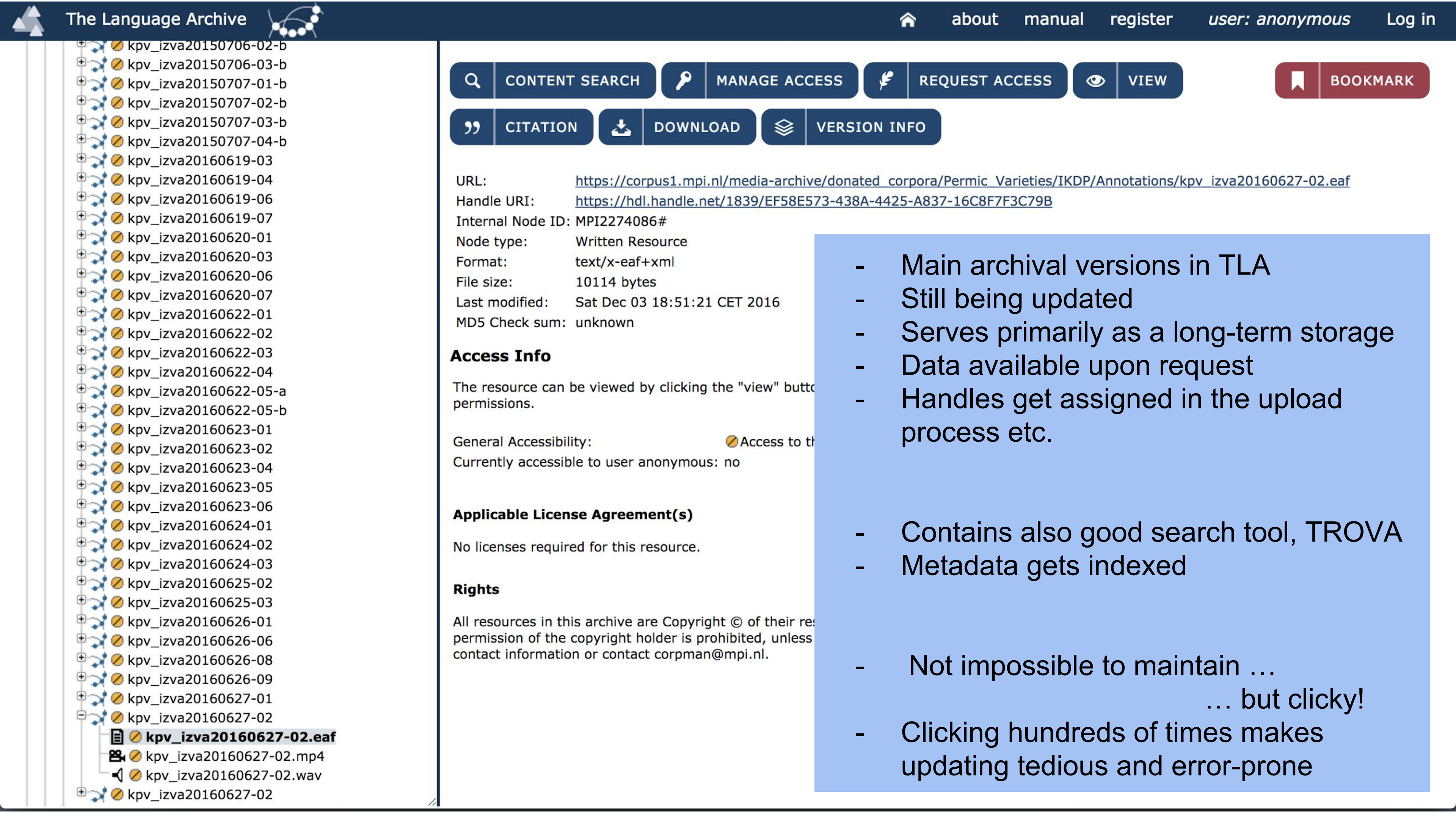

FINLAND
SAINT PETERSBURG
LATVIA
UANIA
MOSCOW
BELARUS
NOVO
Leaflet | © OpenStreetMap © CartoDB

Data availability

- Close collaboration with local institutions and speakers in Russia
- Emphasis in easily shared, rather neutral, still generally interesting topics
  - Biographies, local history, experiences on language use and learning, narratives etc.
  - Central point has been to produce data that in principle can be shared
  - However, with personal data situation is never that simple

- Different users have different needs
  - One interface cannot satisfy all, but some combination probably goes far
  - Raw data access essential so people can do what they want

- Archive is the most important component in the infrastructure we use
  - Nothing else comes even close with long term data preservation



kpv_izva20150706-02-b
kpv_izva20150706-03-b
kpv_izva20150707-01-b
kpv_izva20150707-02-b
kpv_izva20150707-03-b
kpv_izva20150707-04-b
kpv_izva20160619-03
kpv_izva20160619-04
kpv_izva20160619-06
kpv_izva20160619-07
kpv_izva20160620-01
kpv_izva20160620-03
kpv_izva20160620-06
kpv_izva20160620-07
kpv_izva20160622-01
kpv_izva20160622-02
kpv_izva20160622-03
kpv_izva20160622-04
kpv_izva20160622-05-a
kpv_izva20160622-05-b
kpv_izva20160623-01
kpv_izva20160623-02
kpv_izva20160623-04
kpv_izva20160623-05
kpv_izva20160623-06
kpv_izva20160624-01
kpv_izva20160624-02
kpv_izva20160624-03
kpv_izva20160625-02
kpv_izva20160625-03
kpv_izva20160626-01
kpv_izva20160626-06
kpv_izva20160626-08
kpv_izva20160626-09
kpv_izva20160627-01
kpv_izva20160627-02
  kpv_izva20160627-02.eaf
  kpv_izva20160627-02.mp4
  kpv_izva20160627-02.wav
kpv_izva20160627-02

URL:	https://corpus1.mpi.nl/media-archive/donated_corpora/Permic_Varieties/IKDP/Annotations/kpv_izva20160627-02.eaf
Handle URI:	https://hdl.handle.net/1839/EF58E573-438A-4425-A837-16C8F7F3C79B
Internal Node ID:	MPI2274086#
Node type:	Written Resource
Format:	text/x-eaf+xml
File size:	10114 bytes
Last modified:	Sat Dec 03 18:51:21 CET 2016
MD5 Check sum:	unknown

Access Info

The resource can be viewed by clicking the "view" butto
permissions.

General Accessibility: Access to th
Currently accessible to user anonymous: no

Applicable License Agreement(s)

No licenses required for this resource.

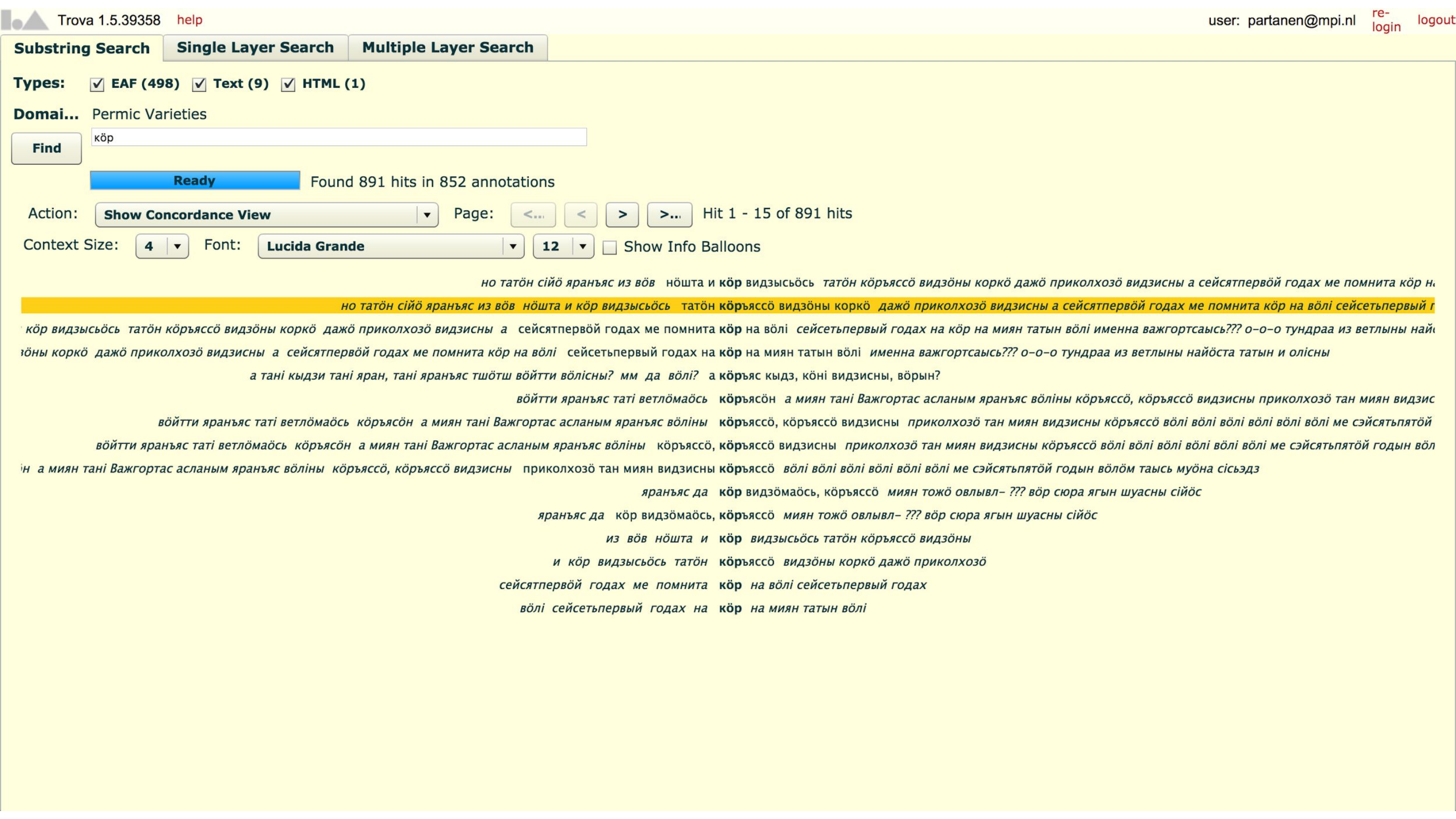
Rights

All resources in this archive are Copyright © of their res
permission of the copyright holder is prohibited, unless
contact information or contact corpman@mpi.nl.

- Main archival versions in TLA
- Still being updated
- Serves primarily as a long-term storage
- Data available upon request
- Handles get assigned in the upload process etc.

- Contains also good search tool, TROVA
- Metadata gets indexed

- Not impossible to maintain …
  … but clicky!
- Clicking hundreds of times makes updating tedious and error-prone



Types: ☑ EAF (498) ☑ Text (9) ☑ HTML (1)

Domai... Permic Varieties

кöр

Ready Found 891 hits in 852 annotations

Action: Show Concordance View

Page: Hit 1 - 15 of 891 hits

Context Size: 4

Font: Lucida Grande 12

☐ Show Info Balloons

но татён сійö яраньяс из вöв нöшта и **кöр** видзысьöсь *татён кöръяссö видзöны коркö дажö приколхозö видзисны а сейсятпервöй годах ме помнита кöр н*

но татён сійö яраньяс из вöв нöшта и кöр видзысьöсь татён **кöр**ъяссö видзöны коркö *дажö приколхозö видзисны а сейсятпервöй годах ме помнита кöр на вöлі сейсетьпервый г*

кöр видзысьöсь татён кöръяссö видзöны коркö дажö приколхозö видзисны а сейсятпервöй годах ме помнита **кöр** на вöлі *сейсетьпервый годах на кöр на миян татын вöлі именна важгортсаысь??? о-о-о тундраа из ветлыны най*

зöны коркö дажö приколхозö видзисны а сейсятпервöй годах ме помнита кöр на вöлі сейсетьпервый годах на **кöр** на миян татын вöлі *именна важгортсаысь??? о-о-о тундраа из ветлыны найöста татын и олісны*

а тані кыдзи тані яран, тані яраньяс тшöтш вöйтти вöлісны? мм да вöлі? а **кöр**ъяс кыдз, кöні видзисны, вöрын?

вöйтти яраньяс таті ветлöмаöсь **кöр**ъясöн *а миян тані Важгортас асланым яраньяс вöліны кöръяссö, кöръяссö видзисны приколхозö тан миян видзис*

вöйтти яраньяс таті ветлöмаöсь кöръясöн а миян тані Важгортас асланым яраньяс вöліны **кöр**ъяссö, кöръяссö видзисны *приколхозö тан миян видзисны кöръяссö вöлі вöлі вöлі вöлі вöлі вöлі ме сэйсятьпятöй*

вöйтти яраньяс таті ветлöмаöсь кöръясöн а миян тані Важгортас асланым яраньяс вöліны кöръяссö, **кöр**ъяссö видзисны *приколхозö тан миян видзисны кöръяссö вöлі вöлі вöлі вöлі вöлі вöлі ме сэйсятьпятöй годын вöл*

н а миян тані Важгортас асланым яраньяс вöліны кöръяссö, кöръяссö видзисны приколхозö тан миян видзисны **кöр**ъяссö *вöлі вöлі вöлі вöлі вöлі вöлі ме сэйсятьпятöй годын вöлöм таысь муöна сісьэдз*

яраньяс да **кöр** видзöмаöсь, кöръяссö *миян тожö овлывл– ??? вöр сюра ягын шуасны сійöс*

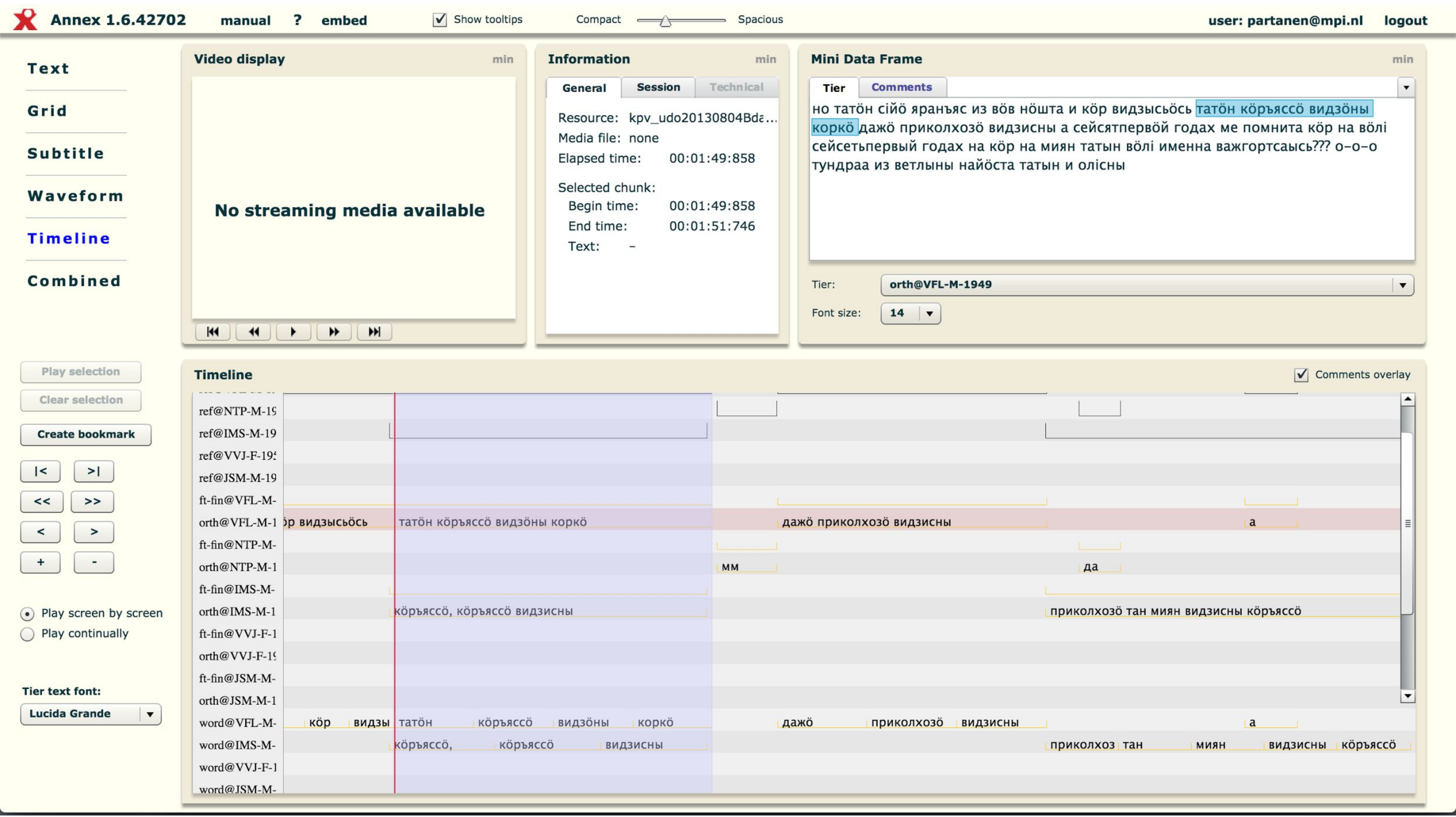
яраньяс да кöр видзöмаöсь, **кöр**ъяссö *миян тожö овлывл– ??? вöр сюра ягын шуасны сійöс*

из вöв нöшта и **кöр** *видзысьöсь татён кöръяссö видзöны*

и кöр видзысьöсь татён **кöр**ъяссö *видзöны коркö дажö приколхозö*

сейсятпервöй годах ме помнита **кöр** *на вöлі сейсетьпервый годах*

вöлі сейсетьпервый годах на **кöр** *на миян татын вöлі*



Annex 1.6.42702
manual
?
embed
☑ Show tooltips
Compact
Spacious
user: partanen@mpi.nl
logout
Text
Grid
Subtitle
Waveform
Timeline
Combined
Video display
min
No streaming media available
Information
min
General
Session
Technical
Mini Data Frame
min
Tier
Comments
Play selection
Clear selection
Create bookmark
|<
>|
<<
>>
<
>
+
-
● Play screen by screen
○ Play continually
Tier text font:
Lucida Grande
Timeline
☑ Comments overlay

Resource: kpv_udo20130804Bdz…
Media file: none
Elapsed time: 00:01:49:858

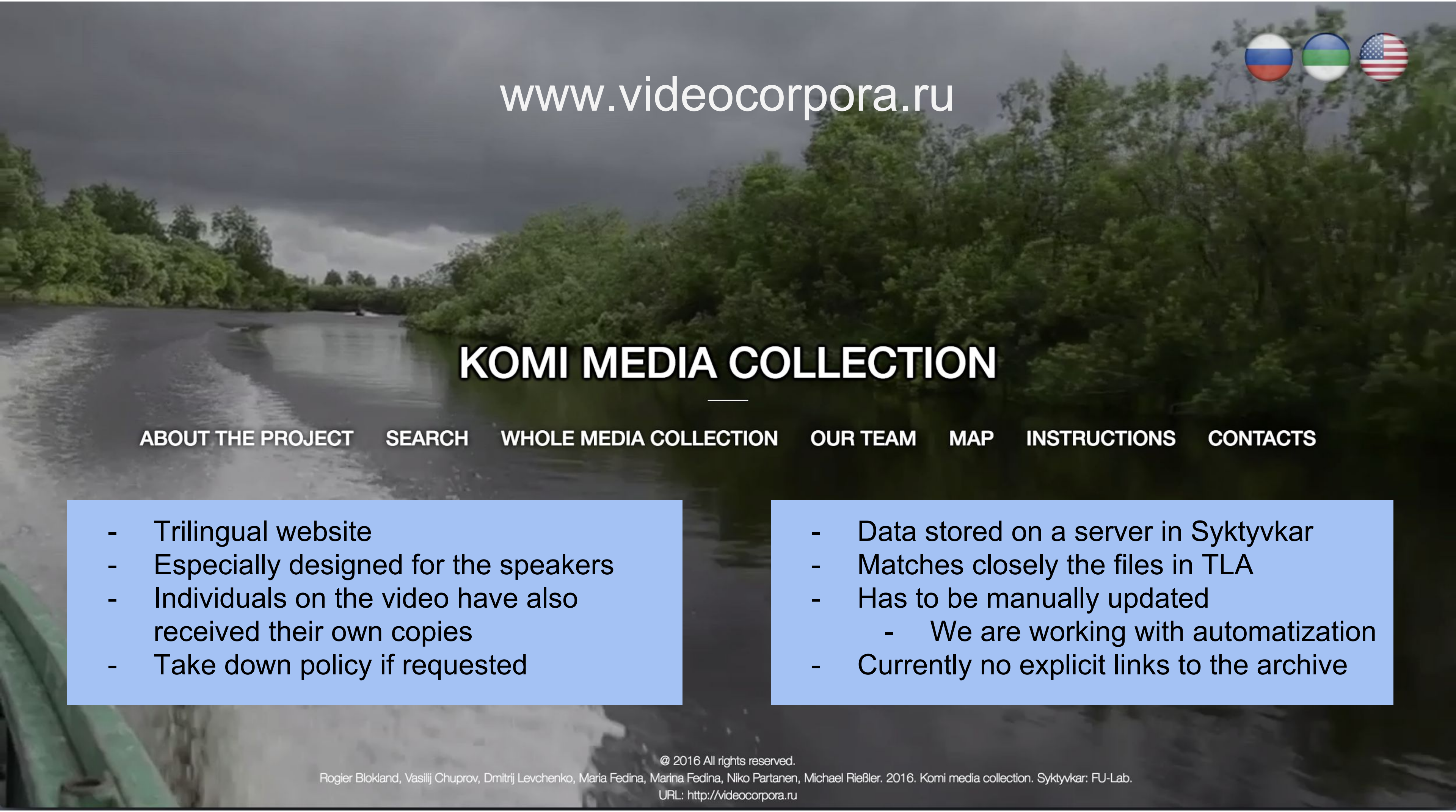
Selected chunk:
Begin time: 00:01:49:858
End time: 00:01:51:746
Text: –

но татӧн сійӧ яранъяс из вӧв нӧшта и кӧр видзысьӧсь татӧн кӧръяссӧ видзӧны коркӧ дажӧ приколхозӧ видзисны а сейсятпервӧй годах ме помнита кӧр на вӧлі сейсетьпервый годах на кӧр на миян татын вӧлі именна важгортсаысь??? о-о-о тундраа из ветлыны найӧста татын и олiсны

Tier: orth@VFL-M-1949
Font size: 14

ref@NTP-M-19
ref@IMS-M-19
ref@VVJ-F-19
ref@JSM-M-19
ft-fin@VFL-M-
orth@VFL-M-1 ӧр видзысьӧсь | татӧн кӧръяссӧ видзӧны коркӧ | дажӧ приколхозӧ видзисны | а
ft-fin@NTP-M-
orth@NTP-M-1 мм | да
ft-fin@IMS-M-
orth@IMS-M-1 кӧръяссӧ, кӧръяссӧ видзисны | приколхозӧ тан миян видзисны кӧръяссӧ
ft-fin@VVJ-F-1
orth@VVJ-F-19
ft-fin@JSM-M-
orth@JSM-M-1
word@VFL-M- кӧр | видзы | татӧн | кӧръяссӧ | видзӧны | коркӧ | дажӧ | приколхозӧ | видзисны | а
word@IMS-M- кӧръяссӧ, | кӧръяссӧ | видзисны | приколхоз | тан | миян | видзисны | кӧръяссӧ
word@VVJ-F-1
word@JSM-M-

www.videocorpora.ru



KOMI MEDIA COLLECTION

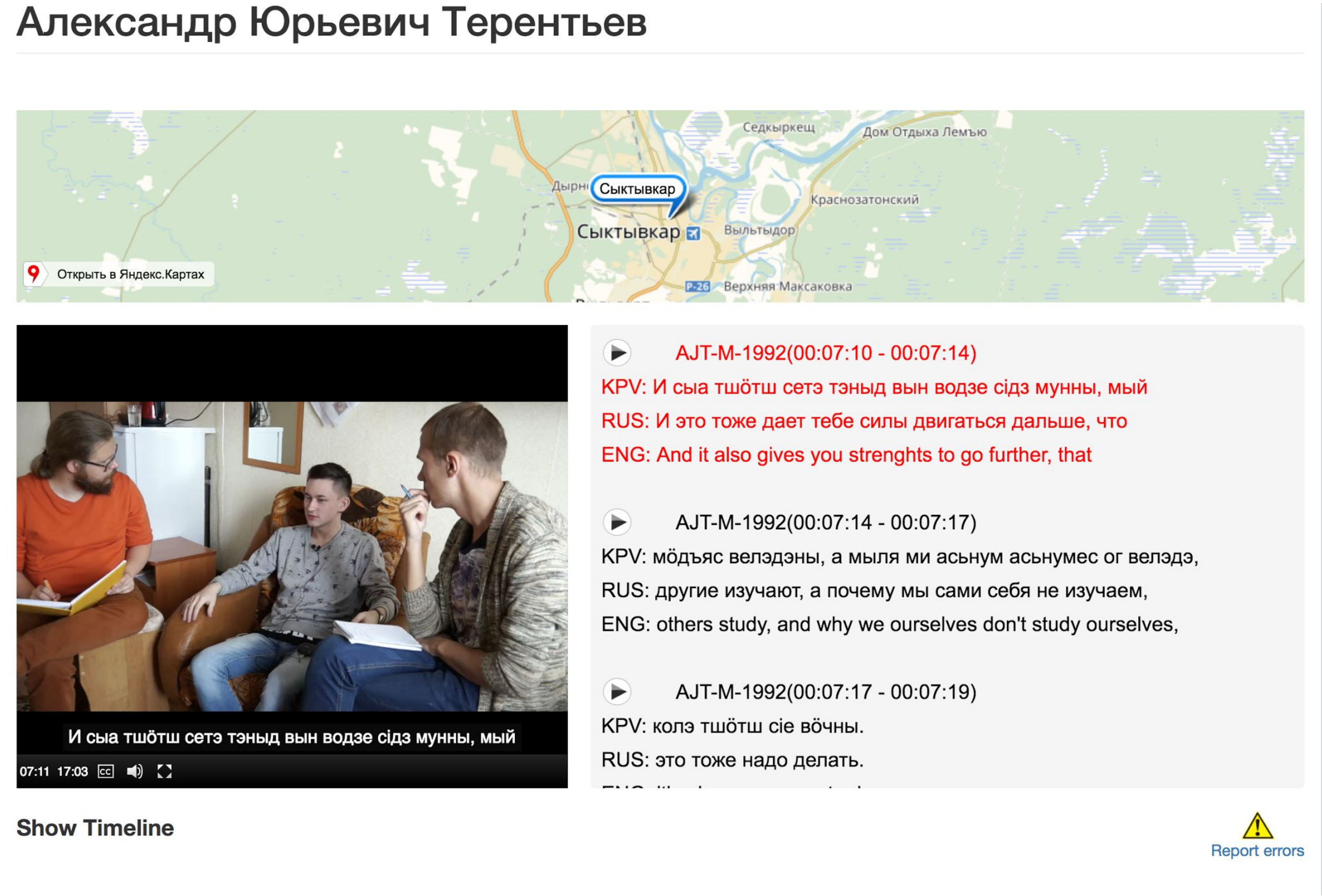
ABOUT THE PROJECT
SEARCH
WHOLE MEDIA COLLECTION
OUR TEAM
MAP
INSTRUCTIONS
CONTACTS

- Trilingual website
- Especially designed for the speakers
- Individuals on the video have also received their own copies
- Take down policy if requested

- Data stored on a server in Syktyvkar
- Matches closely the files in TLA
- Has to be manually updated
    - We are working with automatization
- Currently no explicit links to the archive

@ 2016 All rights reserved.

Rogier Blokland, Vasilij Chuprov, Dmitrij Levchenko, Maria Fedina, Marina Fedina, Niko Partanen, Michael Rießler. 2016. Komi media collection. Syktyvkar: FU-Lab.
URL: http://videocorpora.ru



Александр Юрьевич Терентьев

Седкыркещ
Дом Отдыха Лемъю
Дырні
Сыктывкар
Краснозатонский
Сыктывкар
Выльтыдор
Открыть в Яндекс.Картах
Р-26
Верхняя Максаковка
07:11 17:03

И сыа тшӧтш сетэ тэныд вын водзе сiдз мунны, мый

AJT-M-1992(00:07:10 - 00:07:14)

KPV: И сыа тшӧтш сетэ тэныд вын водзе сiдз мунны, мый

RUS: И это тоже дает тебе силы двигаться дальше, что

ENG: And it also gives you strenghts to go further, that

AJT-M-1992(00:07:14 - 00:07:17)

KPV: мӧдъяс велэдэны, а мыля ми асьнум асьнумес ог велэдэ,

RUS: другие изучают, а почему мы сами себя не изучаем,

ENG: others study, and why we ourselves don't study ourselves,

AJT-M-1992(00:07:17 - 00:07:19)

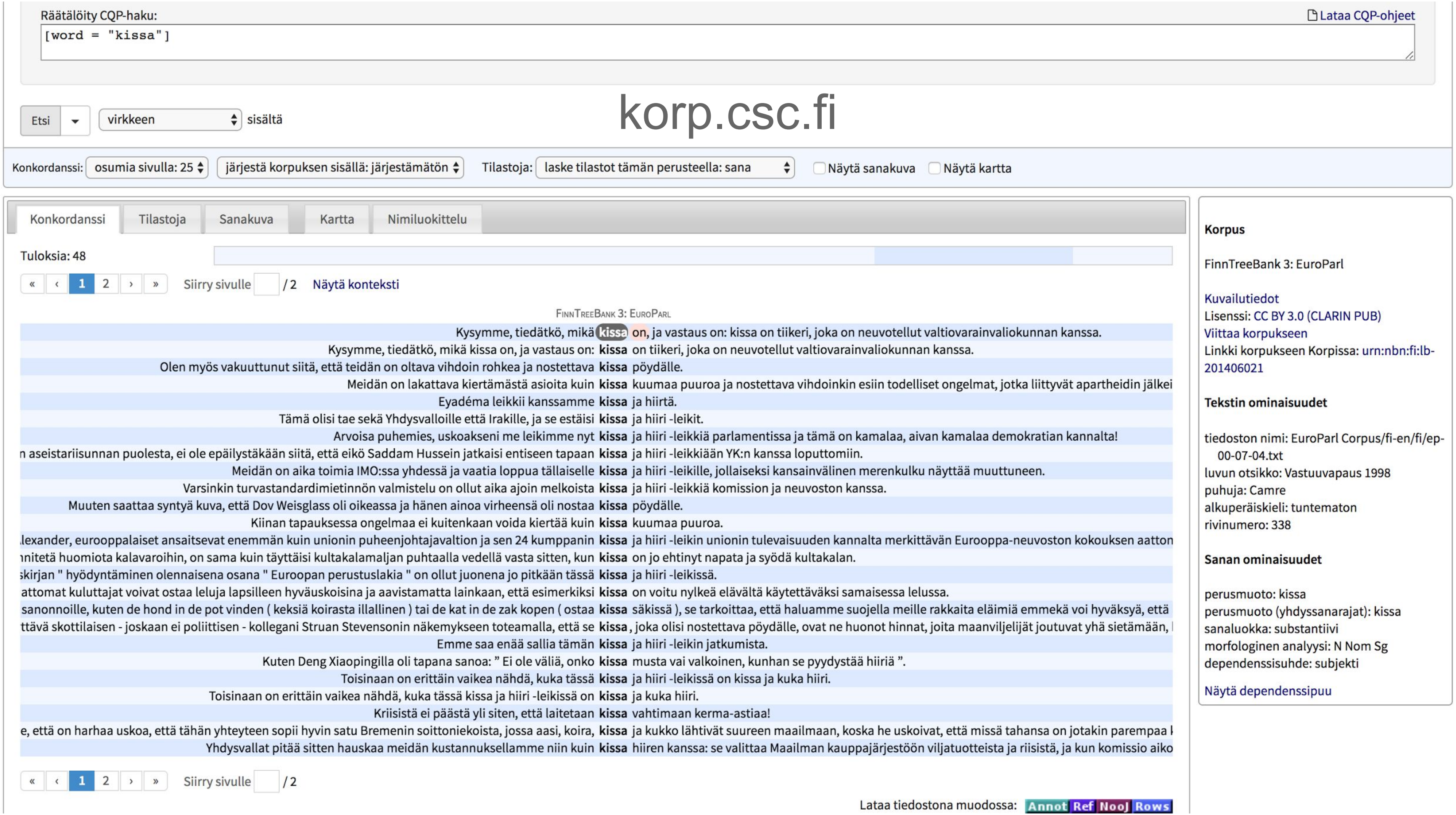
KPV: колэ тшӧтш сiе вӧчны.

RUS: это тоже надо делать.

Show Timeline

Report errors

korp.csc.fi



```
[word = "kissa"]
```

Etsi
virkkeen
sisältä
Konkordanssi: osumia sivulla: 25
järjestä korpuksen sisällä: järjestämätön
Tilastoja: laske tilastot tämän perusteella: sana
☐ Näytä sanakuva
☐ Näytä kartta
Konkordanssi
Tilastoja
Sanakuva
Kartta
Nimiluokittelu

Tuloksia: 48

FinnTreeBank 3: EuroParl

Kysymme, tiedätkö, mikä	**kissa**	on, ja vastaus on: kissa on tiikeri, joka on neuvotellut valtiovarainvaliokunnan kanssa.
Kysymme, tiedätkö, mikä kissa on, ja vastaus on:	**kissa**	on tiikeri, joka on neuvotellut valtiovarainvaliokunnan kanssa.
Olen myös vakuuttunut siitä, että teidän on oltava vihdoin rohkea ja nostettava	**kissa**	pöydälle.
Meidän on lakattava kiertämästä asioita kuin	**kissa**	kuumaa puuroa ja nostettava vihdoinkin esiin todelliset ongelmat, jotka liittyvät apartheidin jälkei
Eyadéma leikkii kanssamme	**kissa**	ja hiirtä.
Tämä olisi tae sekä Yhdysvalloille että Irakille, ja se estäisi	**kissa**	ja hiiri -leikit.
Arvoisa puhemies, uskoakseni me leikimme nyt	**kissa**	ja hiiri -leikkiä parlamentissa ja tämä on kamalaa, aivan kamalaa demokratian kannalta!
n aseistariisunnan puolesta, ei ole epäilystäkään siitä, että eikö Saddam Hussein jatkaisi entiseen tapaan	**kissa**	ja hiiri -leikkiään YK:n kanssa loputtomiin.
Meidän on aika toimia IMO:ssa yhdessä ja vaatia loppua tällaiselle	**kissa**	ja hiiri -leikille, jollaiseksi kansainvälinen merenkulku näyttää muuttuneen.
Varsinkin turvastandardimietinnön valmistelu on ollut aika ajoin melkoista	**kissa**	ja hiiri -leikkiä komission ja neuvoston kanssa.
Muuten saattaa syntyä kuva, että Dov Weisglass oli oikeassa ja hänen ainoa virheensä oli nostaa	**kissa**	pöydälle.
Kiinan tapauksessa ongelmaa ei kuitenkaan voida kiertää kuin	**kissa**	kuumaa puuroa.
lexander, eurooppalaiset ansaitsevat enemmän kuin unionin puheenjohtajavaltion ja sen 24 kumppanin	**kissa**	ja hiiri -leikin unionin tulevaisuuden kannalta merkittävän Eurooppa-neuvoston kokouksen aatton
nnitetä huomiota kalavaroihin, on sama kuin täyttäisi kultakalamaljan puhtaalla vedellä vasta sitten, kun	**kissa**	on jo ehtinyt napata ja syödä kultakalan.
skirjan " hyödyntäminen olennaisena osana " Euroopan perustuslakia " on ollut juonena jo pitkään tässä	**kissa**	ja hiiri -leikissä.
attomat kuluttajat voivat ostaa leluja lapsilleen hyväuskoisina ja aavistamatta lainkaan, että esimerkiksi	**kissa**	on voitu nylkeä elävältä käytettäväksi samaisessa lelussa.
sanonnoille, kuten de hond in de pot vinden (keksiä koirasta illallinen) tai de kat in de zak kopen (ostaa	**kissa**	säkissä), se tarkoittaa, että haluamme suojella meille rakkaita eläimiä emmekä voi hyväksyä, että
ttävä skottilaisen - joskaan ei poliittisen - kollegani Struan Stevensonin näkemykseen toteamalla, että se	**kissa**	, joka olisi nostettava pöydälle, ovat ne huonot hinnat, joita maanviljelijät joutuvat yhä sietämään,
Emme saa enää sallia tämän	**kissa**	ja hiiri -leikin jatkumista.
Kuten Deng Xiaopingilla oli tapana sanoa: ” Ei ole väliä, onko	**kissa**	musta vai valkoinen, kunhan se pyydystää hiiriä ”.
Toisinaan on erittäin vaikea nähdä, kuka tässä	**kissa**	ja hiiri -leikissä on kissa ja kuka hiiri.
Toisinaan on erittäin vaikea nähdä, kuka tässä kissa ja hiiri -leikissä on	**kissa**	ja kuka hiiri.
Kriisistä ei päästä yli siten, että laitetaan	**kissa**	vahtimaan kerma-astiaa!
e, että on harhaa uskoa, että tähän yhteyteen sopii hyvin satu Bremenin soittoniekoista, jossa aasi, koira,	**kissa**	ja kukko lähtivät suureen maailmaan, koska he uskoivat, että missä tahansa on jotakin parempaa
Yhdysvallat pitää sitten hauskaa meidän kustannuksellamme niin kuin	**kissa**	hiiren kanssa: se valittaa Maailman kauppajärjestöön viljatuotteista ja riisistä, ja kun komissio aiko

Korpus

FinnTreeBank 3: EuroParl

Kuvailutiedot
Lisenssi: CC BY 3.0 (CLARIN PUB)
Viittaa korpukseen
Linkki korpukseen Korpissa: urn:nbn:fi:lb-201406021

Tekstin ominaisuudet

tiedoston nimi: EuroParl Corpus/fi-en/fi/ep-00-07-04.txt
luvun otsikko: Vastuuvapaus 1998
puhuja: Camre
alkuperäiskieli: tuntematon
rivinumero: 338

Sanan ominaisuudet

perusmuoto: kissa
perusmuoto (yhdyssanarajat): kissa
sanaluokka: substantiivi
morfologinen analyysi: N Nom Sg
dependenssisuhde: subjekti

Näytä dependenssipuu

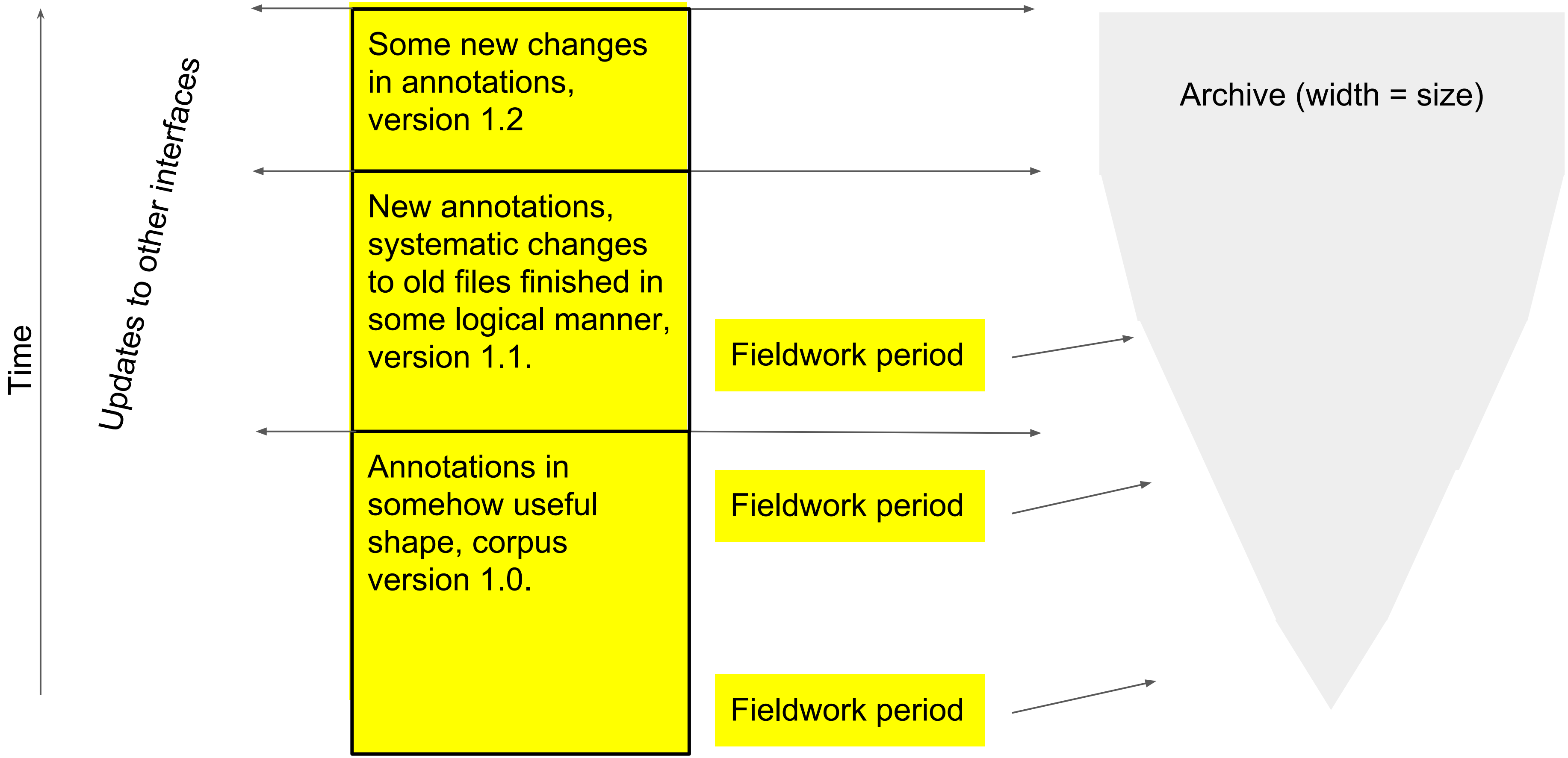
Archive interaction (in an ideal world)

- New audio and video files added roughly once a year
- Once audio and video are archived, the changes into those are exceptional

- There is continuous work going on with annotations
- We use version control (Git) to keep them organized
  - Changes every few days
  - There are usually several larger changes being done

- Ideally the new changed files would be updated in the archive regularly
  - Every half a year?
  - Would correspond to a change in corpus version number

Workflow scheme



Time

Updates to other interfaces

Some new changes in annotations, version 1.2

New annotations, systematic changes to old files finished in some logical manner, version 1.1.

Annotations in somehow useful shape, corpus version 1.0.

Fieldwork period

Fieldwork period

Fieldwork period

Archive (width = size)

Dreaming of API

- At the moment we mainly interact with the archive through specific software
    - Arbil, LAMUS2 etc.
- Works, but is built around the idea that user want to manually do every edit
- We know exactly which files have to be changed and their location in the tree
    - So why do we have to select manually these things again in the interface?
- **Are archives generally even interested about offering APIs to their data?**

- The tasks we need to do are rather modest:
    - Add a file, update a file
    - Change the access level
    - Get information about the files
        - Handles, date of change, MD5, information about number of times accessed etc.

- Data in different interfaces has to be the same!

Thank you!

This work was partially supported by the LATTICE laboratory through a RGNF-CNRS grant.
The research on Iźva Komi has been funded by Kone Foundation.