

KIELIPANKKI
The Language Bank of Finland

Mylly: Uusi tapa käsitellä teksti- ja puheaineistoa helposti ja tehokkaasti

Mietta Lennes ja Jussi Piitulainen
FIN-CLARIN, Helsingin yliopisto

FIN-CLARIN



Kuinka aineistojen käsittelyä voitaisiin helpottaa?

- Aineistojen koko kasvaa jatkuvasti.
- Kaikkia työvaiheita ei pysty tekemään käsin.
- Monet työkalut ovat hankalia käyttää (esim. edellyttävät komentorivityökalujen hallintaa).
- Menetelmät halutaan dokumentoida.
- Työvuon pitäisi olla helposti **toistettavissa** myös uudelle aineistolle.



Chipster

- CSC:n kehittämä avoimen lähdekoodin alusta (<http://chipster.csc.fi/>)
- Käytössä aiemmin luonnontieteiden puolella.



Mylly (The Mill)

- Kielipankkiin on asennettu oma Chipster-versio nimeltään “Mylly”.
- Myllyn kautta tarjotaan erityisesti kieliaineistojen käsittelyyn ja visualisointiin sopivia työkaluja.
- Myllyn kautta on mahdollista kytkeä käyttöön mikä tahansa Taito-palvelimella toimiva työkalu tai skripti, joka käyttää yhtä tai useampaa tiedostoa syötteenä ja tuottaa tulokseksi uusia tiedostoja.



Output = input?



Käyttöesimerkkejä

1. Automaattinen puheentunnistus
2. Tekstin saneistaminen ja sananmuotojen frekvenssit
3. Suomenkielisen tekstin automaattinen jäsentäminen
4. Haku suoraan Korp-palvelusta ja tulokset taulukkomuotoon



KIELIPANKKI

The Language Bank of Finland



KIELIPANKKI KÄYTTÄJÄKSI AINEISTOT TYÖKALUT FOORUMI ORGANISAATIO TUKI

IN ENGLISH PÅ SVENSKA

Myllyn käyttöohjeet

Myllyn on kieliaineiston käsittelyyn ja tutkimiseen tarkoitettu monipuolinen alusta, joka pohjautuu **CSC - Tieteen tietotekniikan keskuksen** kehittämään **Chipster**-teknologiaan. Myllyn kautta voit jo käyttää monia Taito-sovelluspalvelimella olevia työkaluja, esimerkiksi tekstin jäsentimiä ja automaattista puheentunnistinta, ja lisää työkaluja on luvassa.

Myllyn käyttö on helppoa: Kirjaudu CSC:n tunnuksella, lataa käsiteltävä tiedosto palveluun, valitse tarvitsemasi työkalu suoraan valikosta ja paina *Run*. Tulokseksi saamaasi aineistoa voit tarkastella ja jatkokäsitellä muilla Myllyn kytketyillä työkaluilla tai tallentaa omalle koneelle.

Käynnistä Mylly: <http://chipster.csc.fi/mylly.jnlp>

Myllyn kirjaututaan CSC:n myöntämällä käyttäjätunnuksella ja salasanalla.

Lyhyitä esimerkkejä Myllyn käytöstä (PDF, esitelmä Kielitieteen päivillä 2017)

Englanninkielisiä ohjeita



Kuukauden tutkija: Ilmari Ivaska

Uutisia

- Kuukauden tutkija: Ilmari Ivaska (1.5.2017)
- FIN-CLARIN aloittaa Tour de CLARINin (20.4.2017)



Kirjaudutaan Myllyyn

Mylly



Open source platform for
language research data analysis

Initialising Mylly 3.11.6
Connecting to broker at chipster.csc.fi... ok
Logging in...

- ▶ Mylly-testi
- ▶ pohjantuuli
- ▶ pohjantuuli_ja_aurinko.txt
- ▶ pohjantuuli_ja_aurinko.wav
- ▶ query-meta.tsv
- ▶ query-tokens_2.tsv
- ▶ query-tokens.tsv

3:00	16
11:05	422
1:32	104
2:40	684
3:08	17

Login

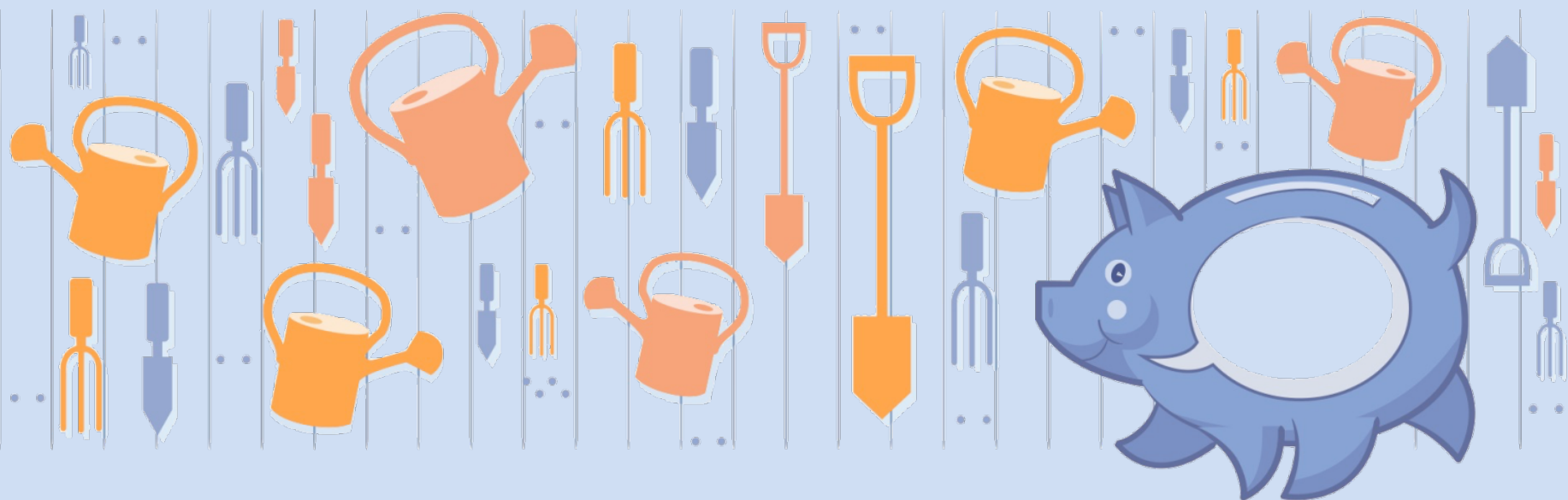
Please enter your Mylly username and password,
or use the username 'guest' and password 'guest'
to have a look. Running tools is disabled for guests.

Username

Password

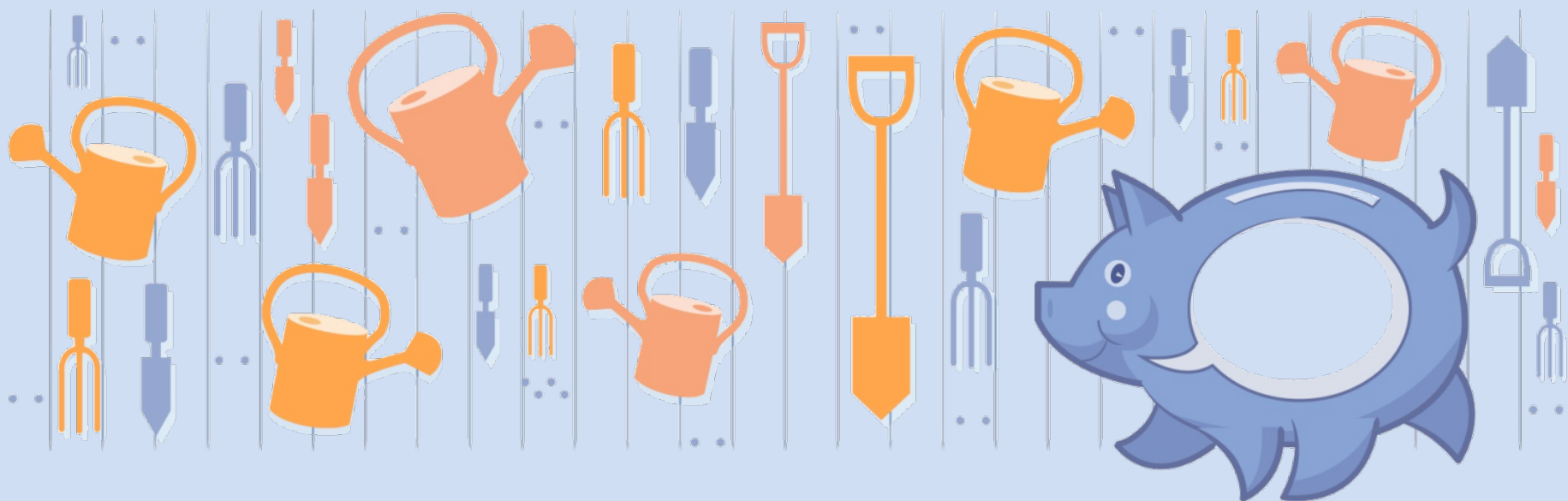


1. Automaattinen puheentunnistus





- **Lähtöaineisto:** Puhetta sisältävä äänitiedosto
- **Työkalu:** Suomen kielen automaattinen puheentunnistin (*AaltoASR*)





Tuodaan Myllyyn äänitiedosto

Mylly 3.10.1

File Edit View Workflow Help

Datasets

To start working with Mylly, you need to load in data first.

[Open example session](#) to get familiar with Mylly

[Open local session](#) to continue working on previous sessions. You can also [open cloud session](#) from the server.

Import new data to Mylly:

- [Import files](#)
- [Import folder](#)
- [Import from URL to client](#)
- [Import from URL directly to server](#)

Analysis tools

Kielipankki

- Korp API
- TSV manipulation
- Syntactic analysis
- Morphological analysis
- Speech recognition**
- Preprocessing
- Finite-State Technology
- Finite-State Transducers
- Job management
- Demo

Aalto Speech Recognizer - Submit Job

Aalto Speech Recognizer - Wait for Results

More help Show tool source

Maximise Detach X

Workflow

Fit

ion

(add your notes)

Aineiston tuonti Myllyyn



Datasets: Valitaan äänitiedosto

Myly 3.10.1

File Edit View Workflow Help

Datasets

- Datasets
 - pohjantuuli_ja_aurinko.wav

Analysis tools

Kielipankki

- Korp API
- TSV manipulation
- Syntactic analysis
- Morphological analysis
- Speech recognition
- Preprocessing
- Finite-State Technology
- Finite-State Transducers
- Job management
- Demo
- Testing

Workflow

Fit

wav

Visualisation

pohjantuuli_ja_aurinko.wav

1 MB, Mon May 15 21:34:38 EEST 2017
(Click here to add your notes)
Created with Chipster 3.10.1
[Analysis hist...](#)

Import / Import data



Analysis tools: Valitaan työkalu

The screenshot shows the Mylly 3.10.1 software interface. The top menu bar includes File, Edit, View, Workflow, and Help. The main window is divided into several panels:

- Datasets:** A list of datasets, with "pohjantuuli_ja_aurinko.wav" selected.
- Analysis tools:** A list of tools including Korp API, TSV manipulation, Syntactic analysis, Morphological analysis, Speech recognition, Preprocessing, Finite-State Technology, Finite-State Transducers, Job management, Demo, and Testing.
- Workflow:** A workspace for building workflows, currently showing a "wav" icon.
- Visualisation:** A panel displaying details for the selected dataset "pohjantuuli_ja_aurinko.wav", including its size (1 MB), creation date (Mon May 15 21:34:38 EEST), and creation tool (Chipster 3.10.1). It also includes a link for "Analysis hist..." and an "Import / Import data" button.

A red speech bubble is overlaid on the right side of the interface, containing the text: **Valitse aineistoon käytettävä työkalu** (Select a tool to use for the dataset).



Lähetetään työ AaltoASR-tunnistimen käsiteltäväksi

Myly 3.10.1

ew Workflow Help

li_ja_aurinko.wav

Analysis tools

Kielipankki

- Korp API
- TSV manipulation
- Syntactic analysis
- Morphological analysis
- **Speech recognition**
- Preprocessing
- Finite-State Technology
- Finite-State Transducers
- Job management
- Demo
- Testing

Aalto Speech Recognizer - Submit Job Show parameters **Run** ▶

Aalto Speech Recognizer - Wait for Results

Submits an audio file for speech recognition in the batch system. "Wait" for the results using the corresponding wait tool on the job file.

More help Show tool sourcecode

Visualisation

Maximise Detach Close

pohjantuuli_ja_aurinko.wav


1 MB, Thu May 18 12:06:08 EEST 2017

(Click here to add your notes)

Created with Chipster 3.10.1

[Analysis hist...](#)

Import / Import data

 [Open in external web browser](#)



(Tunnistustyölle voi ensin valita muutamia asetuksia)

Myly 3.10.1

File Edit View Workflow Help

Datasets

- Datasets
 - pohjantuuli_ja_aurinko.wav

Analysis tools - Speech recognition - Aalto Speech Recognizer - Submit Job

Script	yes	<input checked="" type="checkbox"/> Hide parameters	Run ▶
SegWord	yes	Always output transcript in script.txt	
SegMorph	no		
SegPhone	no		
RawTranscript	yes		

More help Show tool sourcecode

Workflow

Fit

wav

Visualisation

Maximise Detach Close

pohjantuuli_ja_aurinko.wav
1 MB, Mon May 15 21:34:38 EEST 2017
(Click here to add your notes)
Created with Chipster 3.10.1
[Analysis hist...](#)

Import / Import data

[Open in external web browser](#)

Connected to 86.50.168.171

View jobs 0 jobs running Used memory 177M / 800M



Pyydetään tuloksia ja odotetaan...

The screenshot displays the Kielipankki software interface. At the top left, there is a menu bar with 'File', 'Edit', 'View', 'Workflow', and 'Help'. Below the menu bar, there are two panels: 'Datasets' on the left and 'Analysis tools' on the right. The 'Datasets' panel shows a folder named 'Datasets' containing two files: 'pohjantuuli_ja_aurinko.wav' and 'pohjantuuli_ja_aurinko.job'. The 'Analysis tools' panel shows a list of tools under the 'Kielipankki' category, including 'Korp API' and 'TSV manipulation'. In the center, a window titled 'Aalto Speech Recognizer - Submit Job' is open, showing a search bar, a 'Show parameters' button, and a 'Run' button with a green play icon. Below the search bar, there are two tabs: 'Aalto Speech Recognizer - Submit Job' and 'Aalto Speech Recognizer - Wait for Results'. The 'Submit Job' tab is active, and it contains a text area with the following text: 'Waits for the results of a speech recognition job in the batch system. Run selected tool for selected datasets corresponding submit tool.' At the bottom of the window, there are buttons for 'More help' and 'Show tool sourcecode'. In the bottom right corner, there are window control buttons: 'Maximise', 'Detach', and 'Close'. At the bottom left, there is a 'Workflow' panel showing a simple workflow with a 'wav' node connected to a 'job' node.



Tunnistustyö on valmis!

Myly 3.10.1

File Edit View Workflow Help

Datasets

- Datasets
 - pohjantuuli_ja_aurinko.wav
 - pohjantuuli_ja_aurinko.job**
 - pohjantuuli_ja_aurinko.textgrid
 - pohjantuuli_ja_aurinko.eaf
 - stderr.log
 - stdout.log
 - pohjantuuli_ja_aurinko.txt

Analysis tools

Kielipankki

- Korp API
- TSV manipulation
- Syntactic analysis
- Morphological analysis
- Speech recognition**
- Preprocessing
- Finite-State Technology
- Finite-State Transducers
- Job management
- Demo
- Testing

Tool	Description
Aalto Speech Recognizer - Submit Job	Submits a speech recognition job to the batch system.
Aalto Speech Recognizer - Wait for Results	Waits for the results of a speech recognition job in the batch system. The input is the job file from the corresponding submit tool.

Workflow

Fit

```
graph TD; wav[wav] --> job[job]; job --> textg[textg]; job --> eaf[eaf]; job --> log1[log]; job --> log2[log]; job --> txt[txt];
```

Visualisation

Maximise Detach Close

pohjantuuli_ja_aurinko.job

158 B, Mon May 15 21:35:25 EEST 2017
(Click here to add your notes)
Created with Chipster 3.10.1
[Analysis hist...](#)

[Open in external web browser](#)

Speech recognition / Aalto Speech Recognizer - Submit Job

Script	yes
SegWord	yes
SegMorph	no
SegPhone	no
RawTranscript	no

Connected to 86.50.168.171

View jobs 0 jobs running Used memory 107M / 800M



Voidaan katsella tekstimuotoista tulosta...

low
 Fit

```
graph LR; A[ ] --> B[eaf]; A --> C[log]; A --> D[log]; A --> E[txt];
```

Visualisation

Maximise De

pohjantuuli_ja_aurinko.txt
2 kB, Mon May 15 21:38:58 EEST 2017
(Click here to add your notes)
Created with Chipster 3.10.1
[Analysis hist...](#)

Speech recognition / Aalto Speech Recognizer - Wait for Results

[View text](#)
 [Open in exte...](#)



Tunnistettu teksti

Visualisation

View text

Maximise

Detach

Close

Recognizer transcript:

pahin tuuli aurinko pahantuuli ja aurinko väittelivät kummola olisi enemmän voimaa kun he samalla näkivät kulkijan jolla oli yllään lämmin takki n sillä ne sopivat että se on voimakkaampi joka nopeammin saa kulkijan riisumaan takkinsa pohjan tuli alkoi puhaltaa niin että viuhui mutta mitä kovempaa se puhalsi sitä tarkemmin käärimiestäkin ympärille viimein tuli luopua koko hommasta silloin alkoi aurinko loistaa lämpimästi että aikaa kannen kulkija riisui manttelinsa niin oli tuulen pakko myöntää että aurinko oli kuin olikin heistä vahvempi e

Word-level segmentation:

0.656 1.448 pahin
1.464 1.760 tuuli
1.768 2.368 aurinko
2.392 3.560 pahantuuli
3.568 3.656 ja
3.664 4.048 aurinko
4.056 4.640 väittelivät
4.648 5.048 kummola
5.056 5.320 olisi
5.328 5.672 enemmän
5.680 6.160 voimaa

View jobs

0 jobs running

Used memory 115M / 800M



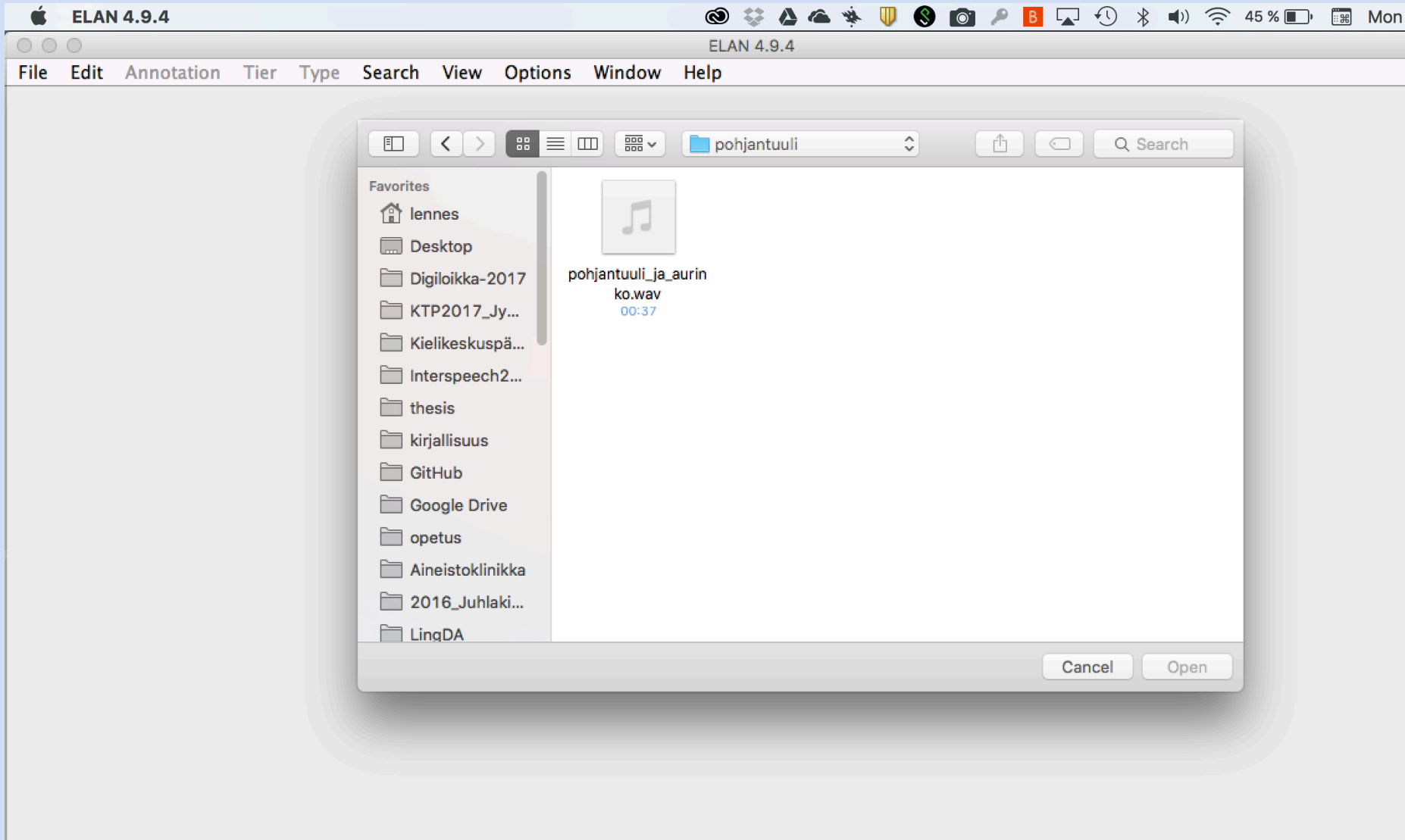
Tunnistustulos EAF- muodossa ELAN-ohjelmaan

The screenshot shows the Mylly 3.10.1 interface with the following components:

- File Menu:** File, Edit, View, Workflow, Help
- Datasets Panel:** Lists files including `pohjantuuli_ja_aurinko.wav`, `pohjantuuli_ja_aurinko.job`, `pohjantuuli_ja_aurinko.textgrid`, `pohjantuuli_ja_aurinko.eaf` (highlighted), `stderr.log`, `stdout.log`, and `pohjantuuli_ja_aurinko.txt`.
- Analysis tools Panel:** Shows a list of tools under the 'Kielipankki' category, including 'Aalto Speech Recognizer - Submit Job' and 'Aalto Speech Recognizer - Wait for Results' (highlighted). Other tools include Korp API, TSV manipulation, Syntactic analysis, Morphological analysis, Speech recognition, Preprocessing, Finite-State Technology, Finite-State Transducers, Job management, Demo, and Testing.
- Workflow Panel:** Displays a workflow diagram where a 'wav' node leads to a 'job' node, which then branches into 'textg', 'eaf', 'log', 'log', and 'txt' nodes. An orange arrow points to the 'eaf' node.
- Visualisation Panel:** Shows details for the file `pohjantuuli_ja_aurinko.eaf`, including its size (26 kB), timestamp (Tue May 16 16:31:54 EEST 2017), and creation tool (Chipster 3.10.1). It also features a link for 'Analysis history' and a button labeled 'Open in external web browser' with an orange arrow pointing to it.

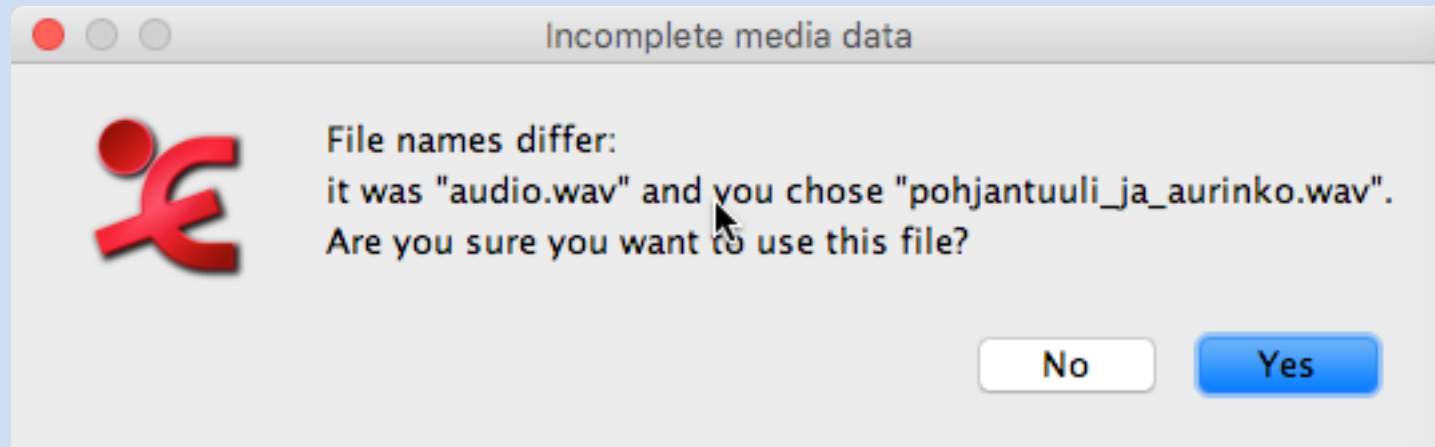


ELAN pyytää valitsemaan vastaavan äänitiedoston omalta koneelta





ELAN ihmettelee vähän, mutta ei välitetä siitä...





Ääni ja annotaatio ELANissa

ELAN 4.9.4

ELAN 4.9.4 - pohjantuuli_ja_aurinko.eaf

File Edit Annotation Tier Type Search View Options Window Help

Grid Text Subtitles Lexicon Comments Recognizers Metadata **Controls**

Volume: 100

pohjantuuli_ja_aurinko.wav

Mute Solo

00:00:06.160 Selection: 00:00:05.680 - 00:00:06.160 480

Selection Mode Loop Mode

pohjantuuli_... 00:00:04.000 00:00:04.500 00:00:05.000 00:00:05.500 00:00:06.000 00:00:06.500 00:00:07.000 00:00:07.500 00:00:08.000

Speaker 1 [79]

väittelivät	kummola	olisi	enemmän	voimaa	kun	he	samalla	näkivät	kulkijan	jol
-------------	---------	-------	---------	--------	-----	----	---------	---------	----------	-----



Tunnistustulos TextGrid- muodossa Praat-ohjelmaan

Myly 3.10.1

File Edit View Workflow Help

Datasets

- Datasets
 - pohjantuuli_ja_aurinko.wav
 - pohjantuuli_ja_aurinko.job
 - pohjantuuli_ja_aurinko.textgrid**
 - pohjantuuli_ja_aurinko.eaf
 - stderr.log
 - stdout.log
 - pohjantuuli_ja_aurinko.txt

Analysis tools

Kielipankki

- Korp API
- TSV manipulation
- Syntactic analysis
- Morphological analysis
- Speech recognition**
- Preprocessing
- Finite-State Technology
- Finite-State Transducers
- Job management
- Demo
- Testing

Aalto Speech Recognizer - Submit Job

Aalto Speech Recognizer - Wait for Results

Show parameters Run

Waits for the results of a speech recognition job from the batch system. The input is the job file from the corresponding submit tool.

More help Show tool source

Workflow

Fit

```
graph TD; wav[wav] --> job[job]; job --> te[te...]; job --> eaf[eaf]; job --> log1[log]; job --> log2[log]; job --> txt[txt];
```

Visualisation

Maximise Detach Close

pohjantuuli_ja_aurinko.textgrid

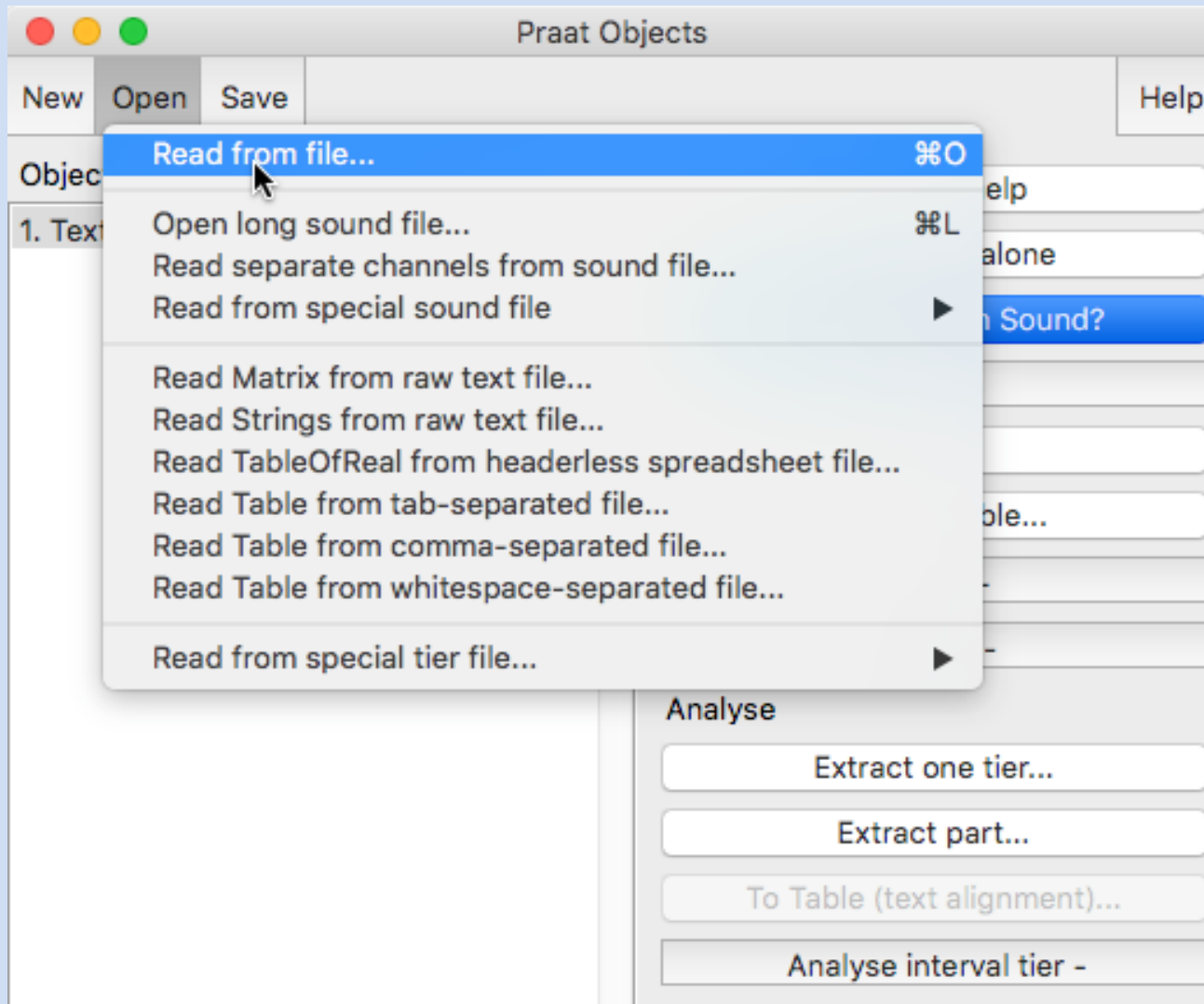
92 kB, Tue May 16 16:31:53 EEST 2017
(Click here to add your notes)
Created with Chipster 3.10.1
[Analysis history](#)

Speech recognition / Aalto Speech Recognizer - Wait for Results

[Open in external web browser](#)

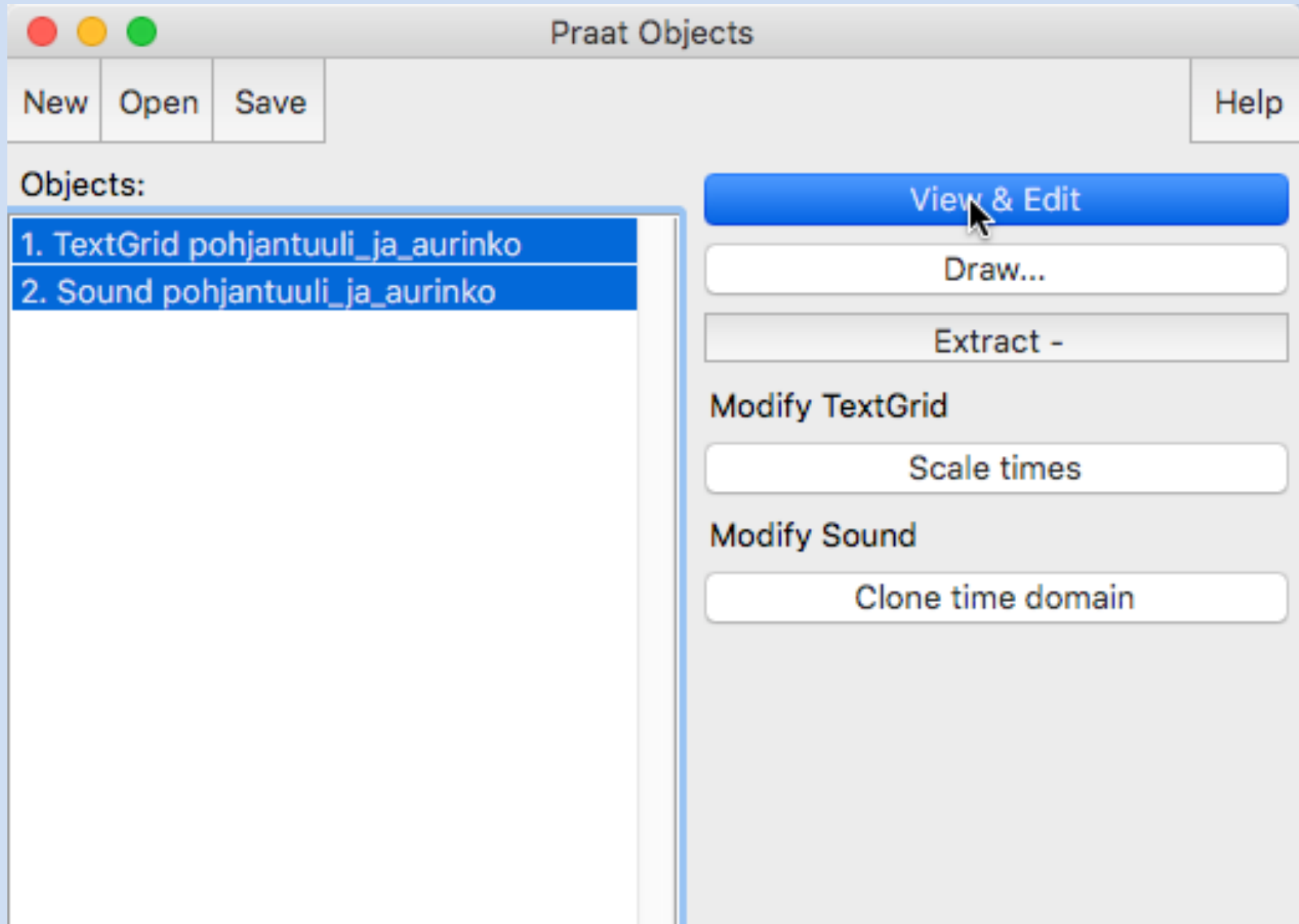


Etsitään omalta koneelta äänitiedosto TextGridin pariksi





Valitaan ääni+TextGrid Praatin editori-ikkunaan





Ääni ja annotaatio Praatissa

1. TextGrid pohjantuuli_ ja_aurinko

File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help

kulkijan

7.608000 0.448000 (2.232 / s) 8.056000

0.3051
0
-0.4193
5000 Hz
2659 Hz
0 Hz

300 Hz
211.5 Hz
120 Hz

1 ku he samalla näkivät **kulkijan** jolla oli yllään lämmin takki word (32/159)

2 ku he sa malla n äki vät kulkija jolla oli yllä än lämmi n takki morph (211)

3 n h e s a m a l l a n ä k i v ä t k u l k i j a j o l l a o l i y l l ä ä l ä m i n t a k phone (546)

0.960333 0.448000 1.956243

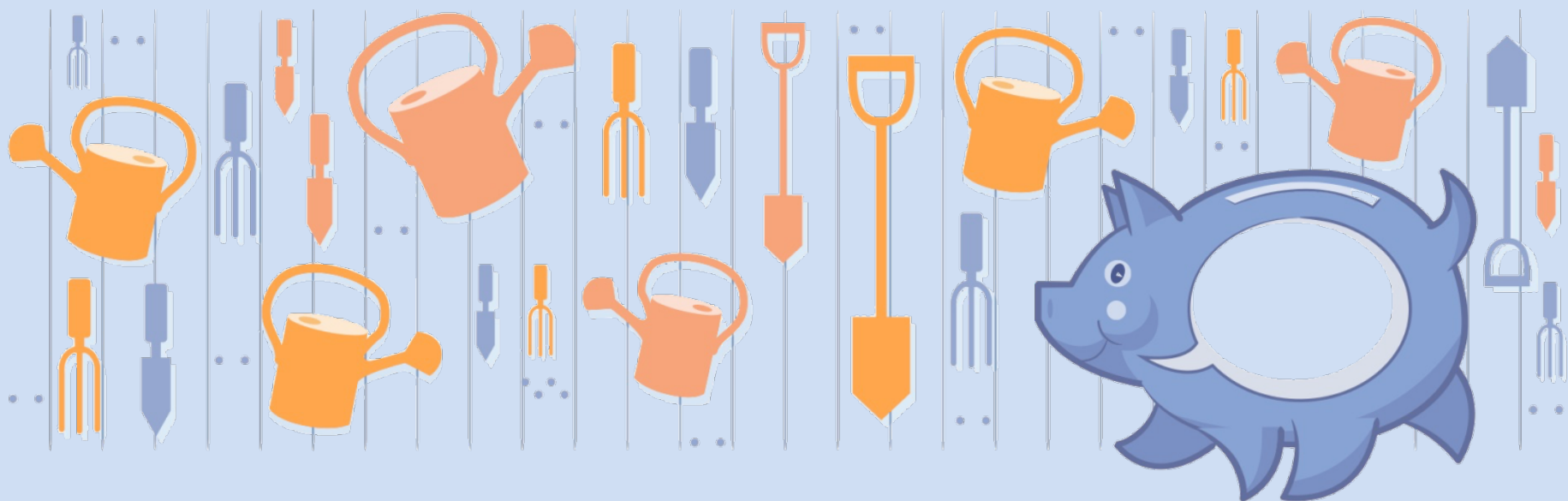
6.647667 6.647667 Visible part 3.364576 seconds 10.012243 26.595757

Total duration 36.608000 seconds

all in out sel bak Group

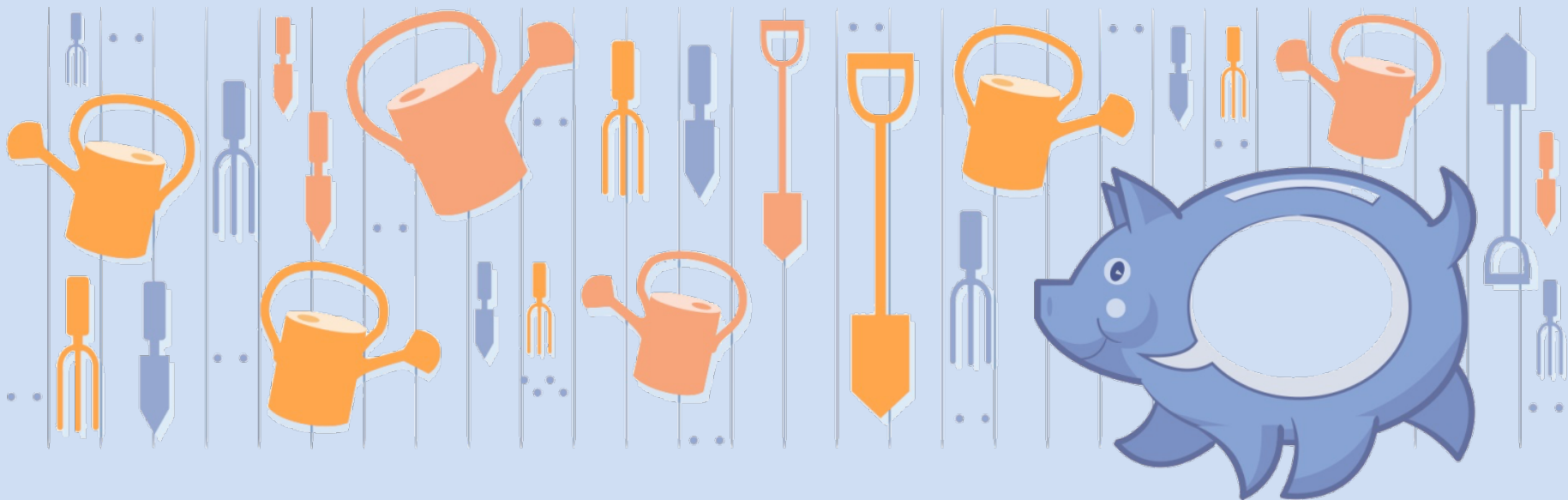


2. Tekstin saneistaminen ja sananmuotojen frekvenssit





- **Lähtöaineisto:** Tekstiä sisältävä tiedosto
- **Työkalu:** Tokenisointi ja sananmuotojen frekvenssien laskenta





Tarvittaessa tuodaan tekstitiedosto Myllyyn

Mylly 3.10.1

File Edit View Workflow Help

Datasets

- Datasets
 - pohjantuuli_ ja_ aurinko.txt

Analysis tools

Kielipankki

- Korp API
- TSV manipulation
- Syntactic analysis
- Morphological analysis
- Speech recognition
- Preprocessing
- Finite-State Technology
- Finite-State Transducers
- Job management
- Demo
- Testing

Power Law Plot of Token Counts
Decreasing Plot of Token Counts
Cumulative Plot of First Occurrences
Location Plot for Selected Words
Simply Tokenize Plain Text
Count Tokens
Trace Tokens

Show parameters Run

Writes one token per line, with line and token number

More help Show tool sourcecode

Workflow

Fit

txt

Visualisation

View text Maximise Detach Close

pahin tuuli aurinko pahantuuli ja aurinko väittelivät kummola olisi enemmän voimaa kun he samalla näkivät kulkijan jolla oli yllään lämmin takki n sillä ne sopivat että se on voimakkaampi joka nopeammin saa kulkijan riisumaan takkinsa pohjan tuli alkoi puhaltaa niin että viuhui mutta mitä kovempaa se puhalsi sitä tarkemmin käärimiestäkin ympärille viimein tuli luopua koko hommasta silloin alkoi aurinko loistaa lämpimästi että aikaa kannen kulkija riisui manttelinsa niin oli tuulen pakko myöntää että aurinko oli kuin olikin heistä vahvempi



Saneistetaan teksti (Tokenize)

Myly 3.10.1

View Workflow Help

ts
pohjantuuli_ja_aurinko.txt

Analysis tools

Kielipankki

- Korp API
- TSV manipulation
- Syntactic analysis
- Morphological analysis
- Speech recognition
- Preprocessing
- Finite-State Technology
- Finite-State Transducers
- Job management
- Demo
- Testing

- Power Law Plot of Token Counts
- Decreasing Plot of Token Counts
- Cumulative Plot of First Occurrences
- Location Plot for Selected Words
- Simply Tokenize Plain Text
- Count Tokens
- Trace Tokens

✓ Show parameters

Writes one token per line, with line number

More help

Show to

v

Fit

Visualisation

☐ Maximise

☐ Detach

pohjantuuli_ja_aurinko.txt

571 B, Tue May 16 16:14:07 EEST 2017

(Click here to add your notes)

Created with Chipster 3.10.1

[Analysis history](#)



[View text](#)



[Open in external web browser](#)



Lasketaan erilaisten sananmuotojen lukumäärät saneistetusta listasta

Myly 3.10.1

File Edit View Workflow Help

Datasets

- Datasets
 - pohjantuuli_ja_aurinko.txt
 - tokens.tsv
 - tokens.txt

Analysis tools

Kielipankki

- Korp API
- TSV manipulation
- Syntactic analysis
- Morphological analysis
- Speech recognition
- Preprocessing
- Finite-State Technology
- Finite-State Transducers
- Job management
- Demo
- Testing

Power Law Plot of Token Counts
Decreasing Plot of Token Counts
Cumulative Plot of First Occurrences
Location Plot for Selected Words
Simply Tokenize Plain Text
Count Tokens
Trace Tokens

Show parameters

Writes word per line, with decreasing order.
Finds token in first field.

More help Show

Workflow

Fit

```
graph TD; txt[txt] --> tsv[tsv]; txt --> txt2[txt];
```

Visualisation

tokens.txt

979 B, Tue May 16 16:14:46 EEST 2017
(Click here to add your notes)
Created with Chipster 3.10.1
[Analysis history](#)

[View text](#)

[Open in external web browser](#)

Demo / Simply Tokenize Plain Text
encoding UTF-8



Mylly 3.10.1

File Edit View Workflow Help

Datasets

- Datasets
 - pohjantuuli_ja_aurinko.txt
 - tokens.tsv
 - tokens.txt
 - countsummary.txt
 - counts.txt**
 - counts.tsv

Analysis tools

Kielipankki

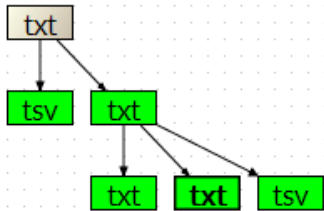
- Korp API
- TSV manipulation
- Syntactic analysis
- Morphological analysis
- Speech recognition
- Preprocessing
- Finite-State Technology
- Finite-State Transducers
- Job management
- Demo**
- Testing

- Power Law Plot of Token Counts
- Decreasing Plot of Token Counts
- Cumulative Plot of First Occurrences
- Location Plot for Selected Words
- Simply Tokenize Plain Text
- Count Tokens**
- Trace Tokens

✓ Sho
Writes wo
Finds toke

Workflow

Fit



Visualisation

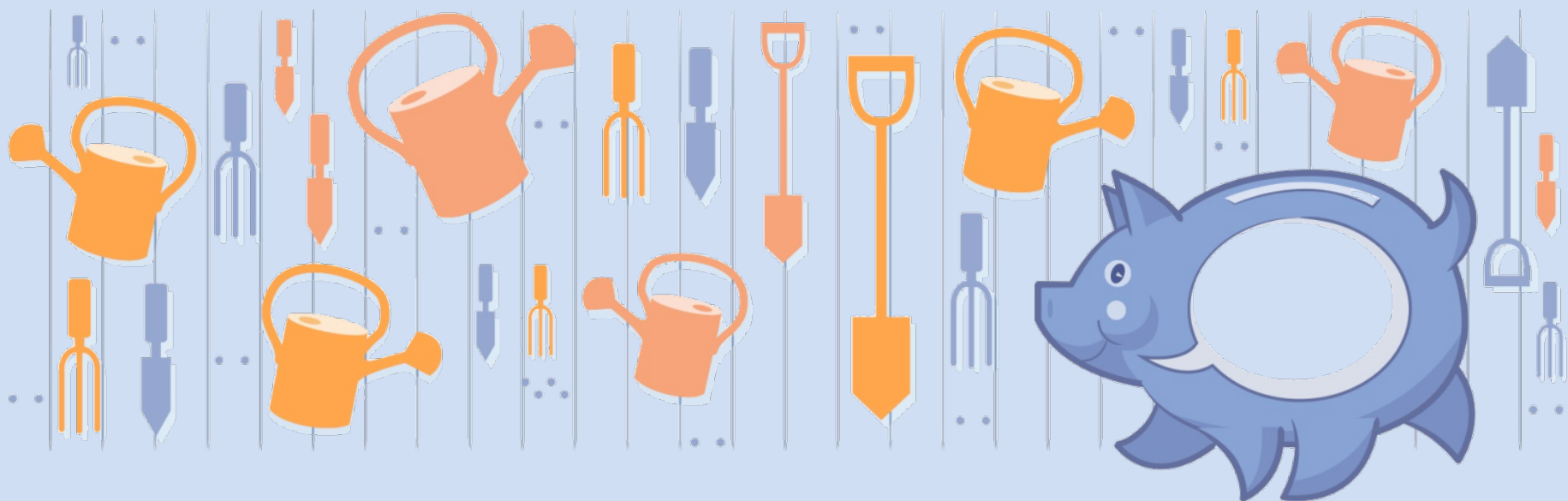
View text

Max

että	4
aurinko	4
oli	3
niin	2
alkoi	2
se	2
tuli	2
kulkijan	2
kovempaa	1
pahin	1
koko	1
kun	1
tuulen	1
lämpimästi	1
takki	1
pohjan	1
aikaa	1
takkinsa	1
käärimestäkin	1
nopeammin	1
sitä	1
n	1
sopivat	1



3. Suomenkielisen tekstin automaattinen jäsentäminen





- **Lähtöaineisto:** Tekstiä sisältävä tiedosto
- **Työkalu:** Suomen kielen dependenssijäsennin
(*Turku Dependency Parser*)





Tekstitiedoston sisällön voi jäsentää Turku Dependency Parser -työkalulla

Mylly 3.10.1

File Edit View Workflow Help

Datasets

- Datasets
 - pohjantuuli_ja_aurinko.txt

Analysis tools

Kielipankki

- Korp API
- TSV manipulation
- Syntactic analysis**
- Morphological analysis
- Speech recognition
- Preprocessing
- Finite-State Technology
- Finite-State Transducers
- Job management
- Demo
- Testing

Turku Dependency Parser for Finnish - Run ✓ Show parameters Run

Turku Dependency Parser for Finnish - Sub

Turku Dependency Parser for Finnish - Wa

Segments Finnish text into sentences and tokens. Annotates each sentence with a morpho-syntactic structure. Runs directly on a server where other people also work.

More help Show tool sourcecode

Workflow

Fit

Visualisation

View text Maximise Detach Close

Pohjantuuli ja aurinko

Pohjantuuli ja aurinko väittelivät kummalla olisi enemmän voimaa, kun he samalla näkivät kulkijan, jolla oli yllään lämmin takki. Silloin he sopivat, että se on voimakkaampi, joka nopeammin saa kulkijan riisumaan takkinsa. Pohjantuuli alkoi puhaltaa niin että viuhui, mutta mitä kovemmin se puhalsi, sitä tiukemmin kääri mies takin ympärilleen, ja viimein tuuli luopui koko hommasta. Silloin alkoi aurinko loistaa lämpimästi, eikä aikaakaan niin kulkija riisui manttelinsa. Niin oli tuulen pakko myöntää, että aurinko oli kuin olikin heistä vahvempi.

Connected to 86.50.168.171

View jobs 0 jobs running Used memory 220M / 800M



Jos jäsennettävä aineisto on iso, työn voi lähettää myös eräajona

Myly 3.10.1

File Edit View Workflow Help

Datasets

- Datasets
 - pohjantuuli_ja_aurinko.txt
 - pohjantuuli_ja_aurinko.job

Analysis tools

Kielipankki		
Korp API	Turku Dependency Parser for Finnish - Run Directly	
TSV manipulation	Turku Dependency Parser for Finnish - Submit Job	
Syntactic analysis	Turku Dependency Parser for Finnish - Wait for Results	Waits for the results of a parsing job in the batch system. The input file is the job file from the corresponding submit tool.
Morphological analysis		
Speech recognition		
Preprocessing		
Finite-State Technology		
Finite-State Transducers		
Job management		
Demo		
Testing		

More help Show tool sourcecode

Workflow

txt
↓
job

Visualisation

Maximise Detach Close

pohjantuuli_ja_aurinko.job
157 B, Tue May 16 15:01:23 EEST 2017
(Click here to add your notes)
Created with Chipster 3.10.1
[Analysis history](#)

[Open in external web browser](#)

Syntactic analysis / Turku Dependency Parser for Finnish - Submit Job

Connected to 86.50.168.171

View jobs 0 jobs running Used memory 105M / 800M



Odotetaan työn valmistumista ja pyydetään tulokset (Wait for results)

Myly 3.10.1

File Edit View Workflow Help

Datasets

- Datasets
 - pohjantuuli_ja_aurinko.txt
 - pohjantuuli_ja_aurinko.job

Analysis tools

Kielipankki		
Korp API	Turku Dependency Parser for Finnish - Run Directly	
TSV manipulation	Turku Dependency Parser for Finnish - Submit Job	
Syntactic analysis	Turku Dependency Parser for Finnish - Wait for Results	Waits for the results of a parsing job in the batch system. The input file is the job file from the corresponding submit tool.
Morphological analysis		
Speech recognition		
Preprocessing		
Finite-State Technology		
Finite-State Transducers		
Job management		
Demo		
Testing		

More help Show tool sourcecode

Workflow

txt
↓
job

Visualisation

pohjantuuli_ja_aurinko.job
157 B, Tue May 16 15:01:23 EEST 2017
(Click here to add your notes)
Created with Chipster 3.10.1
[Analysis history](#)

[Open in external web browser](#)

Syntactic analysis / Turku Dependency Parser for Finnish - Submit Job

Connected to 86.50.168.171

View jobs 0 jobs running Used memory 105M / 800M



TSV-muotoiset tulokset voi avata omalla koneella

Myly 3.10.1

File Edit View Workflow Help

Datasets

- Datasets
 - pohjantuuli_ja_aurinko.txt
 - pohjantuuli_ja_aurinko.job
 - pohjantuuli_ja_aurinko.tsv

Analysis tools

Kielipankki

- Korp API
- TSV manipulation
- Syntactic analysis
- Morphological analysis
- Speech recognition
- Preprocessing
- Finite-State Technology
- Finite-State Transducers
- Job management
- Demo
- Testing

Turku Dependency Parser for Finnish - Run Directly	Shows the results of a parsing job in the batch system. The input file is the job file from the corresponding submit tool.
Turku Dependency Parser for Finnish - Submit Job	
Turku Dependency Parser for Finnish - Wait for Results	

Show parameters Run

More help Show tool sourcecode

Workflow

Fit

```
graph TD; txt[txt] --> job[job]; job --> tsv[tsv];
```

Visualisation

Maximise Detach Close

pohjantuuli_ja_aurinko.tsv

15 rows, Tue May 16 15:04:59 EEST 2017
(Click here to add your notes)
Created with Chipster 3.10.1
[Analysis history](#)

Spreadsheet

Open in external web browser

Syntactic analysis / Turku Dependency Parser for Finnish - Wait for Results

View jobs 0 jobs running Used memory 127M / 800M

Connected to 86.50.168.171



Tulokset TSV-muodossa voivat näyttää tältä oman koneen tekstieditorissa:

pohjantuuli_ja_aurinko.tsv									
1	pahin	paha	A	NUM_Sg CASE_Nom CMP_Superl	2	amod			
2	tuuli	tuuli	N	NUM_Sg CASE_Nom3	nn				
3	aurinko	aurinko	N	NUM_Sg CASE_Nom4	nsubj-cop				
4	pahantuuli	paha tuuli	N	NUM_Sg CASE_Nom7	nsubj				
5	ja	ja	C	SUBCAT_CC	4	cc			
6	aurinko	aurinko	N	NUM_Sg CASE_Nom4	conj				
7	väittelivät	väitellä	V	PRS_Pl3 VOICE_Act TENSE_Prt MOOD_Ind	0		ROOT		
8	kummola	kummola	N	NUM_Sg CASE_Nom OTHER_UNK	11	nsubj-cop			
9	olisi	olla	V	PRS_Sg3 VOICE_Act MOOD_Cond	11	cop			
10	enemmän	enemmän	Adv	-	11	advmod			
11	voimaa	voima	N	NUM_Sg CASE_Par7	dobj				
12	kun	kun	C	SUBCAT_CS	15	mark			
13	he	hän	Pron	SUBCAT_Pers NUM_Pl CASE_Nom	15	nsubj			
14	samalla	samalla	Adv	-	15	advmod			
15	näkevät	nähdä	V	PRS_Pl3 VOICE_Act TENSE_Prt MOOD_Ind	7	advcl			
16	kulkijan	kulkija	N	NUM_Sg CASE_Gen15	dobj				
1	jolla	joka	Pron	SUBCAT_Rel NUM_Sg CASE_Ade	2	rel			
2	oli	olla	V	PRS_Sg3 VOICE_Act TENSE_Prt MOOD_Ind	0		ROOT		
3	yllään	yllä	Adv	POSS_Px3	2	advmod			
1	lämmin	lämmin	A	NUM_Sg CASE_Nom CMP_Pos	2	amod			
2	takki	takki	N	NUM_Sg CASE_Nom3	nn				
3	n	n	N	SUBCAT_Acro NUM_Sg CASE_Nom	0		ROOT		
1	sillä	se	Pron	SUBCAT_Dem NUM_Sg CASE_Ade	3	nommod			
2	ne	se	Pron	SUBCAT_Dem NUM_Pl CASE_Nom	3	nsubj			
3	sopivat	sopia	V	PRS_Pl3 VOICE_Act TENSE_Prt MOOD_Ind	0		ROOT		
4	että	että	C	SUBCAT_CS	7	complm			
5	se	se	Pron	SUBCAT_Dem NUM_Sg CASE_Nom	7	nsubj-cop			
6	on	olla	V	PRS_Sg3 VOICE_Act TENSE_Prs MOOD_Ind	7	cop			
7	voimakkaampi	voimakas	A	NUM_Sg CASE_Nom CMP_Comp	3	ccomp			



TSV-tiedoston voi myös tallentaa Export-komennolla

The screenshot shows the Mylly 3.10.1 application window. The 'Datasets' panel on the left contains a list of files: 'pohjantuuli_ja_aurinko.txt', 'pohjantuuli_ja_aurinko.job', and 'pohjantuuli_ja_aurinko.tsv'. A right-click context menu is open over the 'pohjantuuli_ja_aurinko.tsv' file, with the 'Export...' option highlighted by an orange arrow. The 'Analysis tools' panel on the right lists various tools, with 'Syntactic analysis' selected. The 'Workflow' panel at the bottom left shows a vertical sequence of boxes: 'txt', 'job', and 'tsv'. The 'Visualisation' panel at the bottom right has a dropdown menu set to 'Open in external web browser'.

Valikko tulee näkyviin hiiren oikeaa nappia klikkaamalla!

pohjantuuli_ja_aurinko.tsv is opened in an external web browser

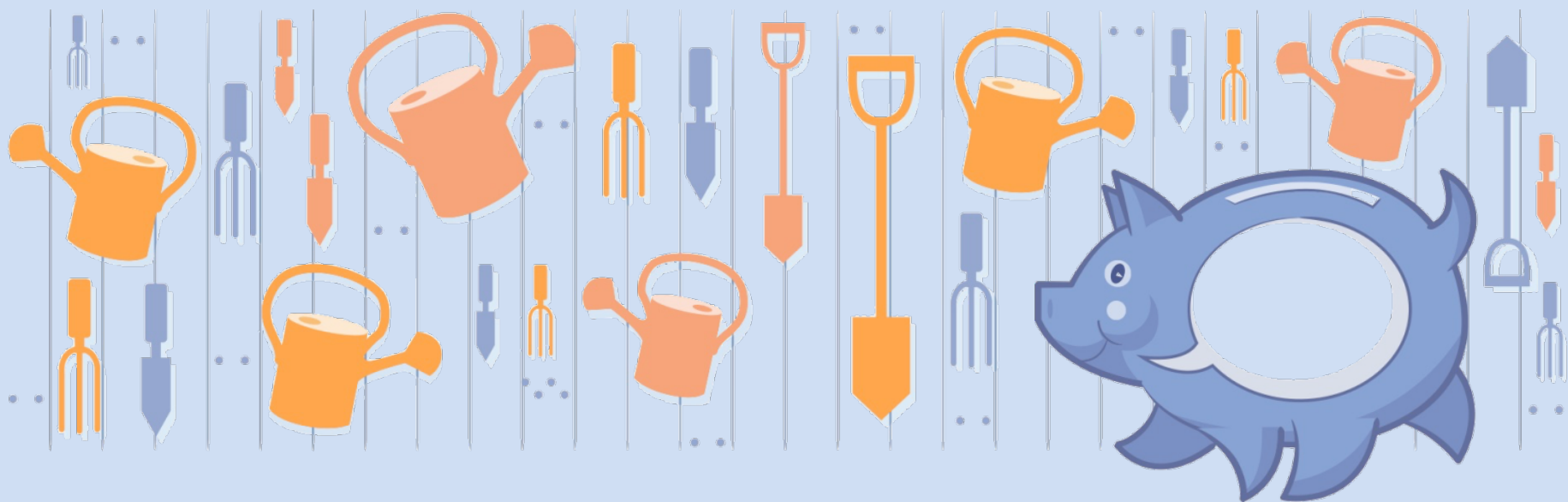


Exceliin tuotuna TSV-tiedosto näyttää tältä:

	A	B	C	D	E	F	G
1	1	pahin	paha	A	NUM_Sg CASE_Nom CMP_Superl	2	amod
2	2	tuuli	tuuli	N	NUM_Sg CASE_Nom	3	nn
3	3	aurinko	aurinko	N	NUM_Sg CASE_Nom	4	nsubj-cop
4	4	pahantuuli	paha tuuli	N	NUM_Sg CASE_Nom	7	nsubj
5	5	ja	ja	C	SUBCAT_CC	4	cc
6	6	aurinko	aurinko	N	NUM_Sg CASE_Nom	4	conj
7	7	väittelivät	väitellä	V	PRS_PI3 VOICE_Act TENSE_Prt MOOD_Ind	0	ROOT
8	8	kummola	kummola	N	NUM_Sg CASE_Nom OTHER_UNK	11	nsubj-cop
9	9	olisi	olla	V	PRS_Sg3 VOICE_Act MOOD_Cond	11	cop
10	10	enemmän	enemmän	Adv	_	11	advmod
11	11	voimaa	voima	N	NUM_Sg CASE_Par	7	dobj
12	12	kun	kun	C	SUBCAT_CS	15	mark
13	13	he	hän	Pron	SUBCAT_Pers NUM_PI CASE_Nom	15	nsubj
14	14	samalla	samalla	Adv	_	15	advmod
15	15	näkevät	nähdä	V	PRS_PI3 VOICE_Act TENSE_Prt MOOD_Ind	7	advcl
16	16	kulkijan	kulkija	N	NUM_Sg CASE_Gen	15	dobj
17							
18	1	iolla	ioka	Pron	SUBCAT_Rel NUM_Sg CASE_Ade	2	rel

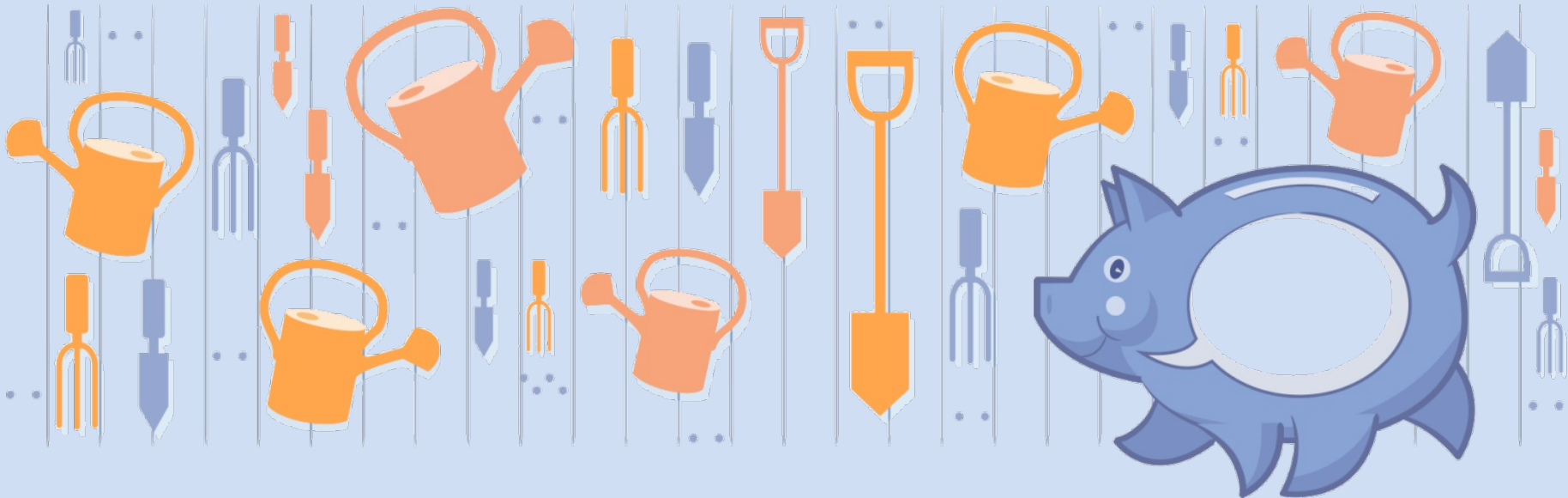


4. Haku suoraan Korp-palvelusta, tulokset taulukkomuodossa





- **Lähtöaineisto:** Korp-palvelun Suomi24 (näyte) -korpus
- **Työkalu:** Korp-API,
<https://www.kielipankki.fi/support/korpapi/>





Luodaan ja tallennetaan ensin tekstitiedosto, jossa on CQP-kyselylause(ita)

```
query.txt
File Path : ~/Desktop/query.txt
query.txt
1 [lemma="kissa"]
2
3 [lemma="koira"]
4
```

Huom! Tiedoston päätteessä **.txt** ja merkistökoodaus **UTF-8!**






Tuodaan kyselytiedosto Myllyyn

Mylly 3.10.1

File Edit View Workflow Help

Datasets

To start working with Mylly, you need to load in data first.

-  [Open example session](#) to get familiar with Mylly
-  [Open local session](#) to continue working on previous sessions. You can also [open cloud session](#) from the server.
-  Import new data to Mylly:
 - [Import files](#)
 - [Import folder](#)
 - [Import from URL to client](#)
 - [Import from URL directly to server](#)

Analysis tools

Kielipankki

- Korp API**
- TSV manipulation
- Syntactic analysis
- Morphological analysis
- Speech recognition
- Preprocessing
- Finite-State Technology
- Finite-State Transducers
- Job management
- Demo
- Testing

KWIC as TSV
[Concordance from Suomi24 corpus in Korp](#)

Workflow

Fit

Visualisation

Unsaved session
(Click here to add your notes)



Valitaan kyselytiedosto ja sitten työkaluksi Korp API

The screenshot shows the Mylly 3.10.1 software interface. The top right corner displays the version number "Mylly 3.10.1". The main menu includes "File", "Edit", "View", "Workflow", and "Help".

The interface is divided into several panels:

- Datasets:** A folder named "Datasets" is expanded, showing a file named "query.txt" which is currently selected.
- Analysis tools:** A list of tools is shown, with "Kielipankki" selected. Under "Kielipankki", "Korp API" is highlighted. Other tools include TSV manipulation, Syntactic analysis, Morphological analysis, Speech recognition, Preprocessing, Finite-State Technology, Finite-State Transducers, Job management, Demo, and Testing.
- Workflow:** A "Fit" checkbox is checked. Below it, a "txt" icon is visible.
- Visualisation:** The file "query.txt" is listed, with a timestamp "23.5.2017, Tue, May 16, 15:40:40 EEST 2017" below it.



Käynnistetään konkordanssihaku

Myly 3.10.1

File Edit View Workflow Help

Datasets

- Datasets
 - query.txt

Analysis tools

Kielipankki		✓ Show parameters	Run ▶
Korp API	KWIC as TSV		
TSV manipulation	Concordance from Suomi24 corpus in Korp		
Syntactic analysis			
Morphological analysis			
Speech recognition			
Preprocessing			
Finite-State Technology			
Finite-State Transducers			
Job management			
Demo			
Testing			

More help Show tool sourcecode

Workflow

txt

Visualisation

query.txt

33 B, Tue May 16 15:40:40 EEST 2017
(Click here to add your notes)
Created with Chipster 3.10.1
[Analysis history](#)

Import / Import data

[View text](#)

[Open in external web browser](#)

Maximise Detach Close

Connected to 86.50.168.171

View jobs 0 jobs running Used memory 186M / 800M



Tuloksena on JSON-muotoinen tekstitiedosto konkordanssista

Mylly 3.10.1

File Edit View Workflow Help

Datasets

- Datasets
 - query.txt
 - query.json

Analysis tools

Kielipankki

- Korp API
 - KWIC as TSV
 - Concordance from Suomi24 corpora
- TSV manipulation
- Syntactic analysis
- Morphological analysis
- Speech recognition
- Preprocessing
- Finite-State Technology
- Finite-State Transducers
- Job management
- Demo
- Testing

Workflow

Fit

```
graph TD; txt[txt] --> json[json];
```

Visualisation

Open in external web browser

query.json is opened in an external



Käynnistetään JSON-muotoisen hakutuloksen muunnos TSV-tiedostoiksi

The screenshot shows the Mylly 3.10.1 software interface. The top menu bar includes File, Edit, View, Workflow, and Help. The main interface is divided into several panels:

- Datasets:** Shows a folder named 'Datasets' containing 'query.txt' and 'query.json'. An orange arrow points to 'query.json'.
- Analysis tools:** A list of tools is shown, with 'Korp API' selected. An orange arrow points to 'Korp API'. The tool 'KWIC as TSV' is highlighted, with an orange arrow pointing to its 'Run' button. The description for 'KWIC as TSV' reads: 'Concordance from Suomi24 corpus in Korp'. Below the description, there are buttons for 'Show parameters', 'Run', 'More help', and 'Show tool sourcecode'.
- Workflow:** Shows a workflow diagram with a 'txt' node pointing to an 'json' node.
- Visualisation:** Displays the output of the 'query.json' tool. It shows the file name 'query.json', its size (4 MB), the date and time (Tue May 16 15:54:26 EEST 2017), and the software version (Created with Chipster 3.10.1). There is a link for 'Analysis history' and a button to 'Open in external web browser'. Below this, the text 'Korp API / Concordance from Suomi24 corpus in Korp' is displayed.



Tuloksena on kaksi TSV-muotoista tekstitiedostoa konkordanssista

Myly 3.10.1

File Edit View Workflow Help

Datasets

- Datasets
 - query.txt
 - query.json
 - query-tokens.tsv
 - query-meta.tsv

Analysis tools

Kielipankki

- Korp API
 - TSV manipulation
 - Syntactic analysis
 - Morphological analysis
 - Speech recognition
 - Preprocessing
 - Finite-State Technology
 - Finite-State Transducers
 - Job management
 - Demo
 - Testing

KWIC as TSV

Concordance from Suomi24 corpus in Korp

✓ 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.

More help Show tool sourcecode

Workflow

Fit

```
graph TD; txt[txt] --> json[json]; json --> tsv1[tsv]; json --> tsv2[tsv];
```

Visualisation

Maximise Detach Close

query.json

4 MB, Tue May 16 15:54:26 EEST 2017

(Click here to add your notes)

Created with Chipster 3.10.1

[Analysis history](#)

[Open in external web browser](#)

Korp API / Concordance from Suomi24 corpus in Korp



TSV-tiedostot voi tallentaa

The screenshot shows the Mylly 3.10.1 application window. The top menu bar includes File, Edit, View, Workflow, and Help. The main interface is divided into several panels:

- Datasets:** A file explorer showing a folder named 'Datasets' containing 'query.txt', 'query.json', 'query-tokens.tsv' (highlighted), and 'query-meta.tsv'. A context menu is open over 'query-tokens.tsv' with options: Visualise, Link to phenodata, Links between selected, Rename, Delete, Save workflow, Export... (highlighted), and View history as text.
- Analysis tools:** A list of tools under the 'Kielipankki' category, including Korp API (selected), TSV manipulation, Syntactic analysis, Morphological analysis, Speech recognition, Preprocessing, Finite-State Technology, Finite-State Transducers, Job management, Demo, and Testing. The 'KWIC as TSV' tool is also visible.
- Workflow:** A diagram showing a flow from 'txt' to 'json' and then to two 'tsv' files.
- Visualisation:** A panel with the option 'Open in external web browser'.

query-tokens.tsv is opened in an external web browser...



TSV-tiedostot voi avata Excelissä (tässä tiedosto *query-tokens.tsv*)

	A	B	C	D	E	F	G	H	I
1	_match	_sen	_tok	dephead	msd	lemma	lex	ref	deprel
2	0	0	0	2	SUBCAT_Rel NUM_Sg CASE_Ine	mikä	mikä..pn.1	1	det
3	0	0	1	3	NUM_PI CASE_Nom	eläin	eläin..nn.1	2	nsubj
4	0	0	2	0	PRS_Sg3 VOICE_Act TENSE_Prs MOOD_Ind	juosta	juosta..vb.1	3	ROOT
5	0	0	3	3	NUM_Sg CASE_Ess CMP_Pos	vapaa	vapaa..jj.1	4	nommod
6	0	0	4	4	_	,	,..xx.1	5	punct
7	0	0	5	4	SUBCAT_CC	ja	ja..kn.1	6	cc
8	0	0	6	9	SUBCAT_Dem NUM_Sg CASE_Nom	se	se..pn.1	7	nsubj-cop
9	0	0	7	9	PRS_Sg3 VOICE_Act TENSE_Prs MOOD_Ind	olla	olla..vb.1	8	cop
10	0	0	8	4	_	sitten	sitten..ab.1	9	conj
11	0	0	9	11	NUM_Sg CASE_Gen	eläin	eläin..nn.1	10	poss
12	0	0	10	13	NUM_Sg CASE_Gen	omistaja	omistaja..nn.1	11	poss
13	0	0	11	13	NUM_Sg CASE_Ade CMP_Pos	oma	oma..jj.1	12	amod
14	0	0	12	9	NUM_Sg CASE_Ade	vastuu	vastuu..nn.1	13	nommod
15	0	0	13	21	_	,	,..xx.1	14	punct
16	0	0	14	21	SUBCAT_CS	jos	jos..kn.1	15	mark
17	0	0	15	18	_	vaikka	vaikka..ab.1	16	advmod
18	0	0	16	18	_	esimerkiksi	esimerkiksi..ab.1	17	advmod
19	1	0	17	21	NUM_Sg CASE_Nom	koira	koira..nn.1	18	nsubj
20	0	0	18	18	SUBCAT_CC	tai	tai..kn.1	19	cc
21	0	0	19	18	NUM_Sg CASE_Nom	kissa	kissa..nn.1	20	conj
22	0	0	20	3	PRS_Sg3 VOICE_Act TENSE_Prs MOOD_Ind	jäää	jäää..vb.1	21	advcl
23	0	0	21	21	NUM_Sg CASE_Gen	auto	auto..nn.1	22	nommod
24	0	0	22	22	SUBCAT_Po	alle	alle..pp.1	23	adpos
25	0	0	23	21	_	,	,..xx.1	24	punct
26	0	0	24	28	SUBCAT_Dem NUM_Sg CASE_Els	se	se..pn.1	25	nommod



TSV-tiedostot voi avata Excelissä (tässä tiedosto *query-meta.tsv*)

	A	B	C	D	E	F	G	H
1	_sen	_start	_end	_corpus	text_time	sentence_id	text_title	text_sub
2	0	17	18	S24	18:27	6061192	Kissan vapaana pito	Yleistä kisso
3	1	24	25	S24	15:04	6062626	Naapurin koira tappanut kissani?	Yleistä kisso
4	2	4	5	S24	23:17	6099514	Kissa viisaampi kuin Koira?	Yleistä kisso
5	3	4	5	S24	12:06	8979264	ISKÄ UHKASI TAPPAA KISSANI JOS (Lävistyksen
6	4	2	3	S24	02:40	6844027	perheenjäsen halvaannutti kanin	Kanit
7	5	1	2	S24	18:38	8100448	Eläinten pito kerrostaloissa pitäisi k	Kerrostalo
8	6	21	22	S24	15:04	6062623	Naapurin koira tappanut kissani?	Yleistä kisso
9	7	15	16	S24	08:06	7634146	Eläimillä ei ole mitään oikeuksia	Eläinten suo
10	8	7	8	S24	02:51	9560104	Mikä rotu sopisi?	Seurakoirat
11	9	14	15	S24	16:40	6138288	Raskaus muutti koirani	Yleistä koiri
12	10	12	13	S24	16:44	1505340	Kissa ja mäyräkoira?	Mäyräkoira
13	11	15	16	S24	07:28	6062853	Naapurin koira tappanut kissani?	Yleistä kisso
14	12	1	2	S24	21:46	6876810	Kertokaa pikkusten nimiä:3	Lemmikkien
15	13	6	7	S24	18:16	508535	Tuholaistorjunta ja lemmikit	Yleistä asun
16	14	3	4	S24	01:17	8169390	Koira hallitsee naisen elämää	Sinkut
17	15	8	9	S24	15:06	1299498	30 tyhmää kysymystä	Gallupit ja k
18	16	4	5	S24	21:38	1333351	Pitkä ajantappogallup	Gallupit ja k
19	17	7	8	S24	11:05	6080501	Voiko kissaa pitää sisällä kerrostalo	Yleistä kisso



Kiitos!

Ehdota meille lisää
Mylly-palveluita:
fin-clarin@helsinki.fi

