Turun yliopiston digitaaliset kieliaineistot ja kieliteknologian työkalut

Monitieteinen esittely- ja demopäivä

UTU-Digilang ja URKO-työryhmä

5.11.2021

TURUN YLIOPISTO

UTU-Digilang

- hyväksytty TY:n tutkimuksen infrastruktuurien tiekartalle vuonna 2021
- kokoaa yhteen TY:n digitaalisia kieliaineistoja ja kieliteknologian työkaluja
  - kehitetty, muodostettu ja ylläpidetty
    - kieli- ja käännöstieteiden laitoksessa
    - tietotekniikan laitoksessa
- siinä yhdistyvät TY:n…
  - pitkät perinteet digitaalisten kieliaineistojen kehittämisessä ja ylläpitämisessä
  - uudet teknologiat ja huippuluokan kieliteknologiaresurssit.

https://sites.utu.fi/utu-digilang/

UTU-Digilang

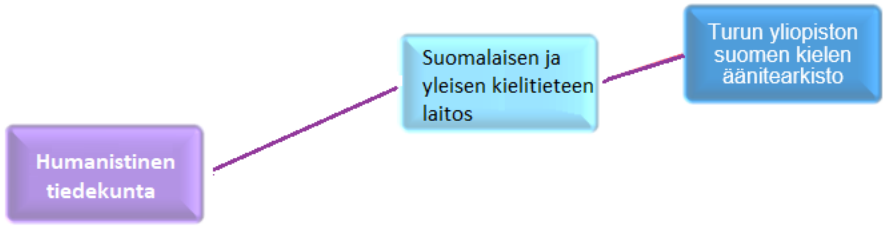
- kokoaa yhteen TY:n digitaalisia kieliaineistoja ja kieliteknologian työkaluja
  - Kieli- ja käännöstieteiden laitos (KKL)
    - Suomen ja sen sukukielten arkisto
      - Lauseopin arkisto (LA)
      - Turun yliopiston suomen kielen äänitearkisto (TYSKÄ)
      - Suomalais-ugrilaiset korpukset
    - Kieli- ja käännöstieteiden laitoksen digitaaliset kieliaineistot ja kieliresurssit
      - Digilang-portaali TY:n digitaaalisten kieliaineistojen hakemiseen
      - Digilang-pitkäaikaistallennus digitaalisten kieliaineistojen keskitettyyn tallennukseen
      - Muita kieliaineita kuin suomea ja suomen sukukieliä koskevat digitaaliset kieliresurssit
  - Tietotekniikan laitoksen ja KKL:n yhteinen TurkuNLP-tutkimusryhmä
    - Turku Neural Parser, FinBERT-kielimalli ja muut kieliteknologiset työkalut
    - Finnish Internet Parsebank ja muut digitaaliset kieliaineistot

Tausta

- Turun yliopistolla yli 50 vuoden perinteet annotoiduista digitaalisista aineistoista
  - Lauseopin arkisto 1967-
- TY:n annotoituja digitaalisia kielikorpuksia/työkaluja/ kieliresursseja
  - Lauseopin arkisto murrekorpus
  - Mormula-korpus
  - Agricola-korpus
  - Arkisyn-korpus
  - Turkulainen suomen kielen puupankki
  - TurkuNER korpus
  - FinBERT

TURUN YLIOPISTO

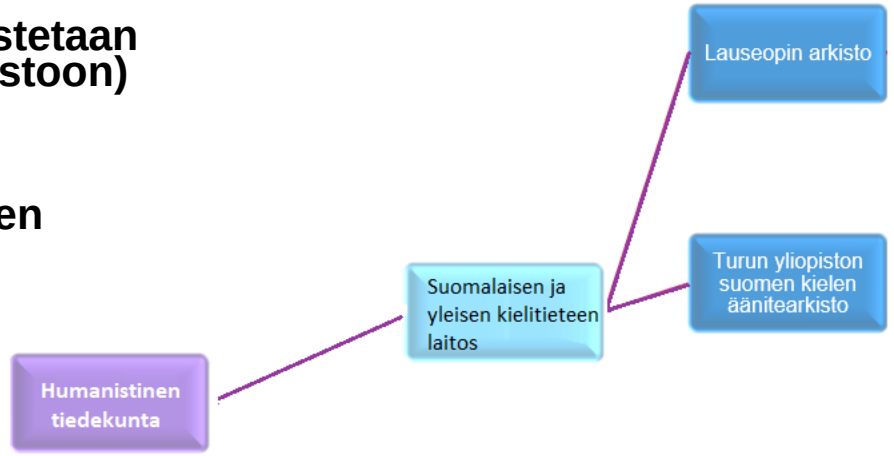
1957



TURUN YLIOPISTO

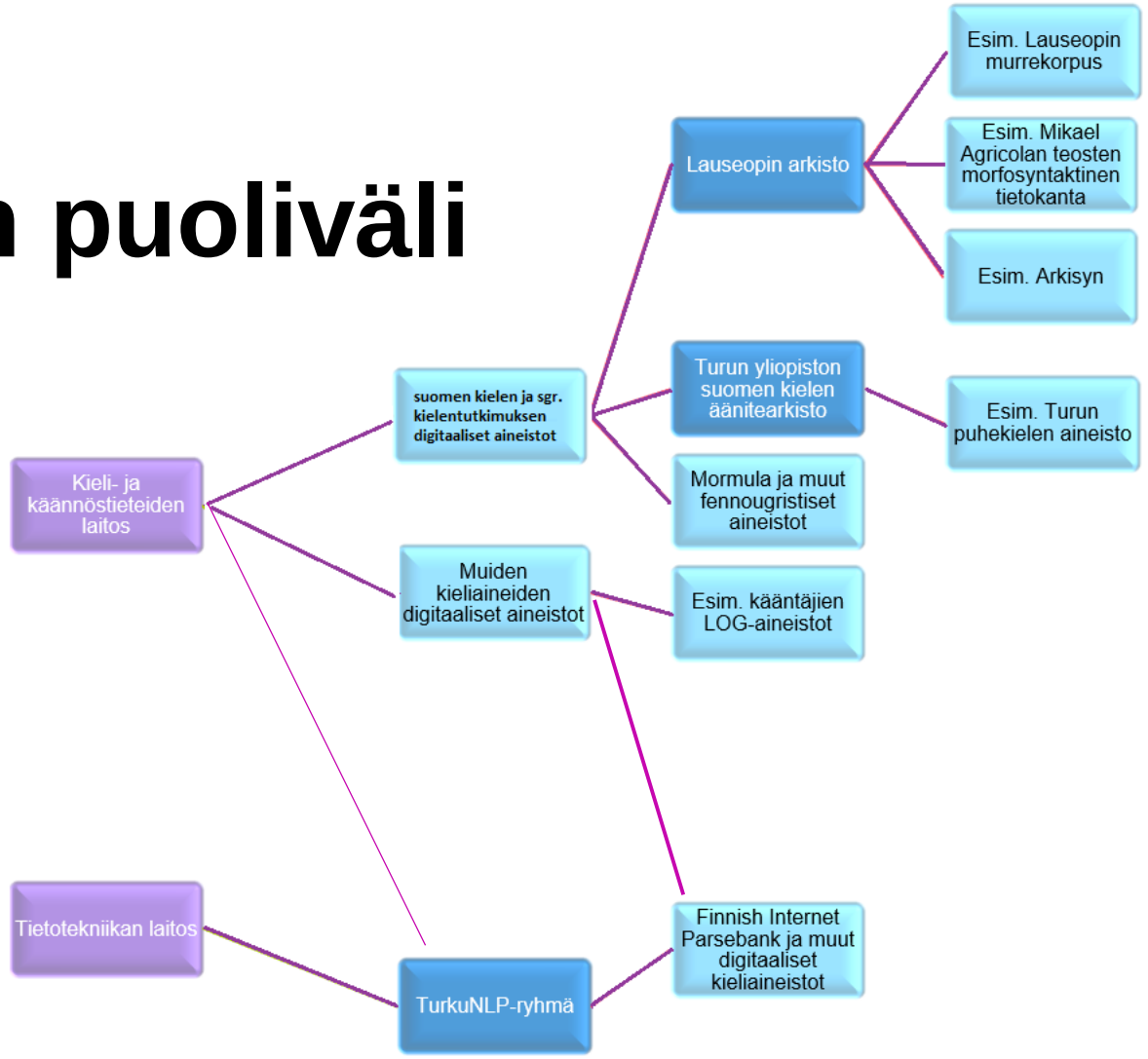
1967

- **Lauseopin arkisto perustetaan (sijoitetaan Turun yliopistoon)**
- **Lauseopin arkiston murrekorpus Suomen ensimmäinen digitaalinen annotoitu kielikorpus**



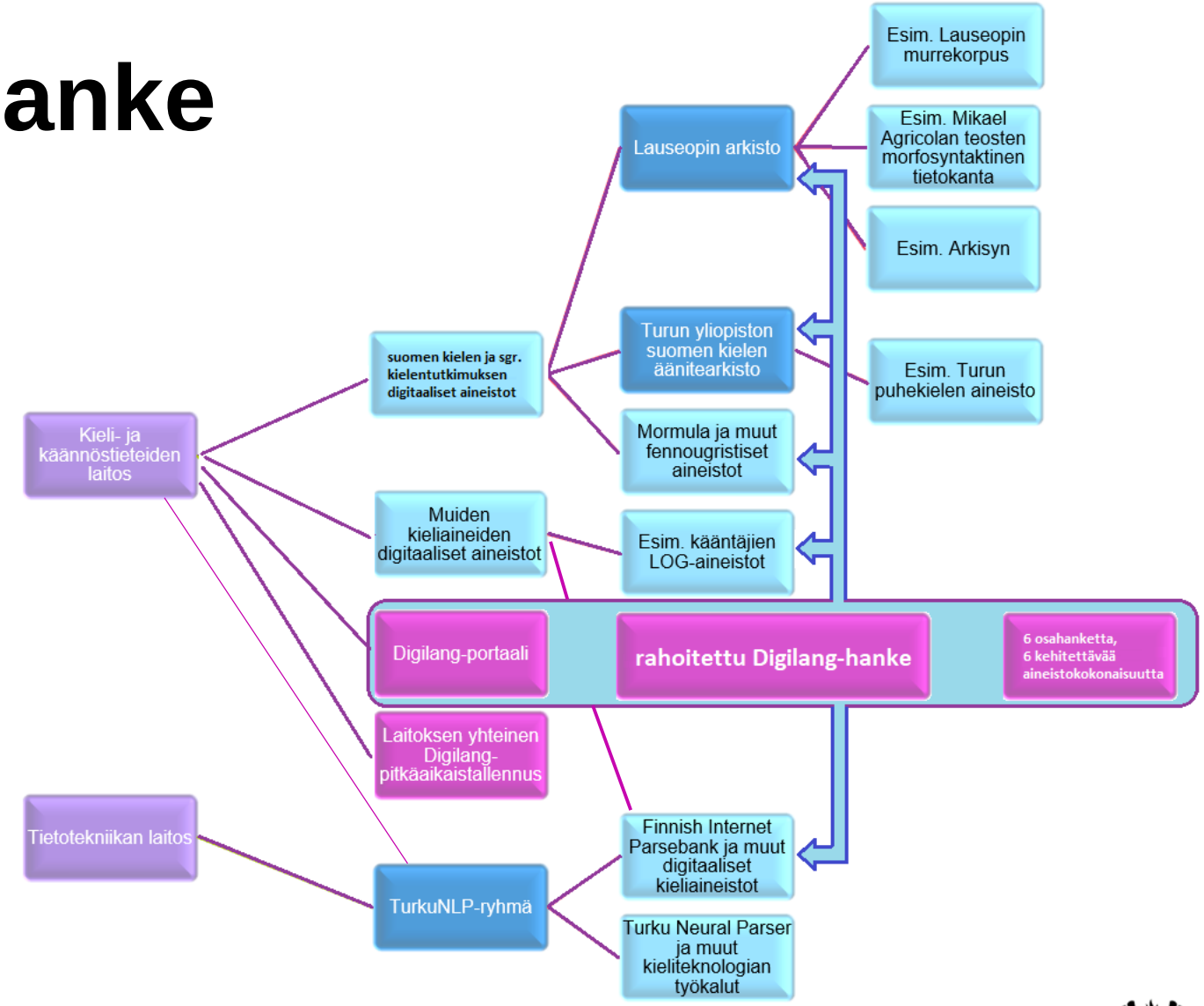
TURUN YLIOPISTO

2010-luvun puoliväli



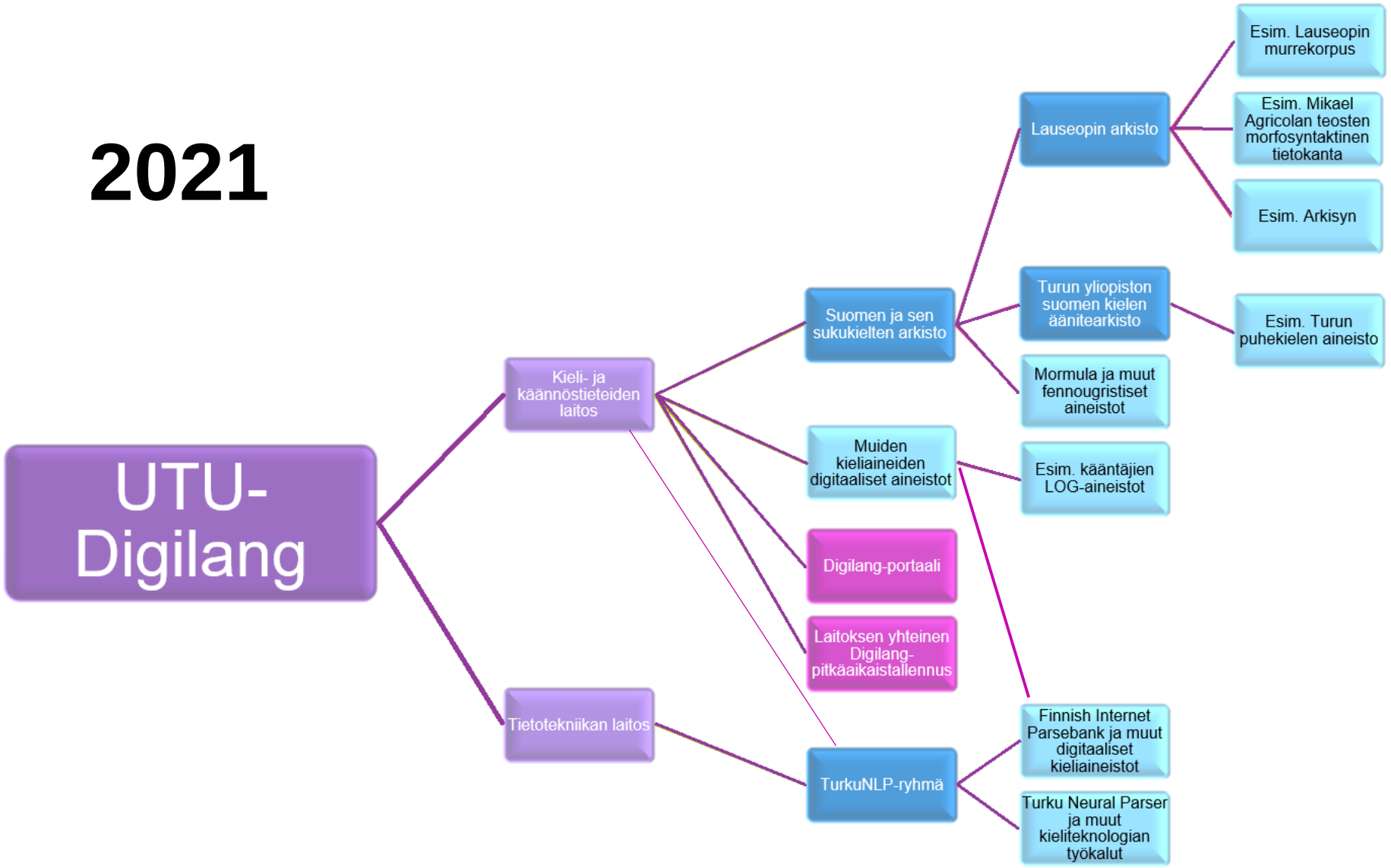
TURUN YLIOPISTO

Digilang-hanke 2018-2021

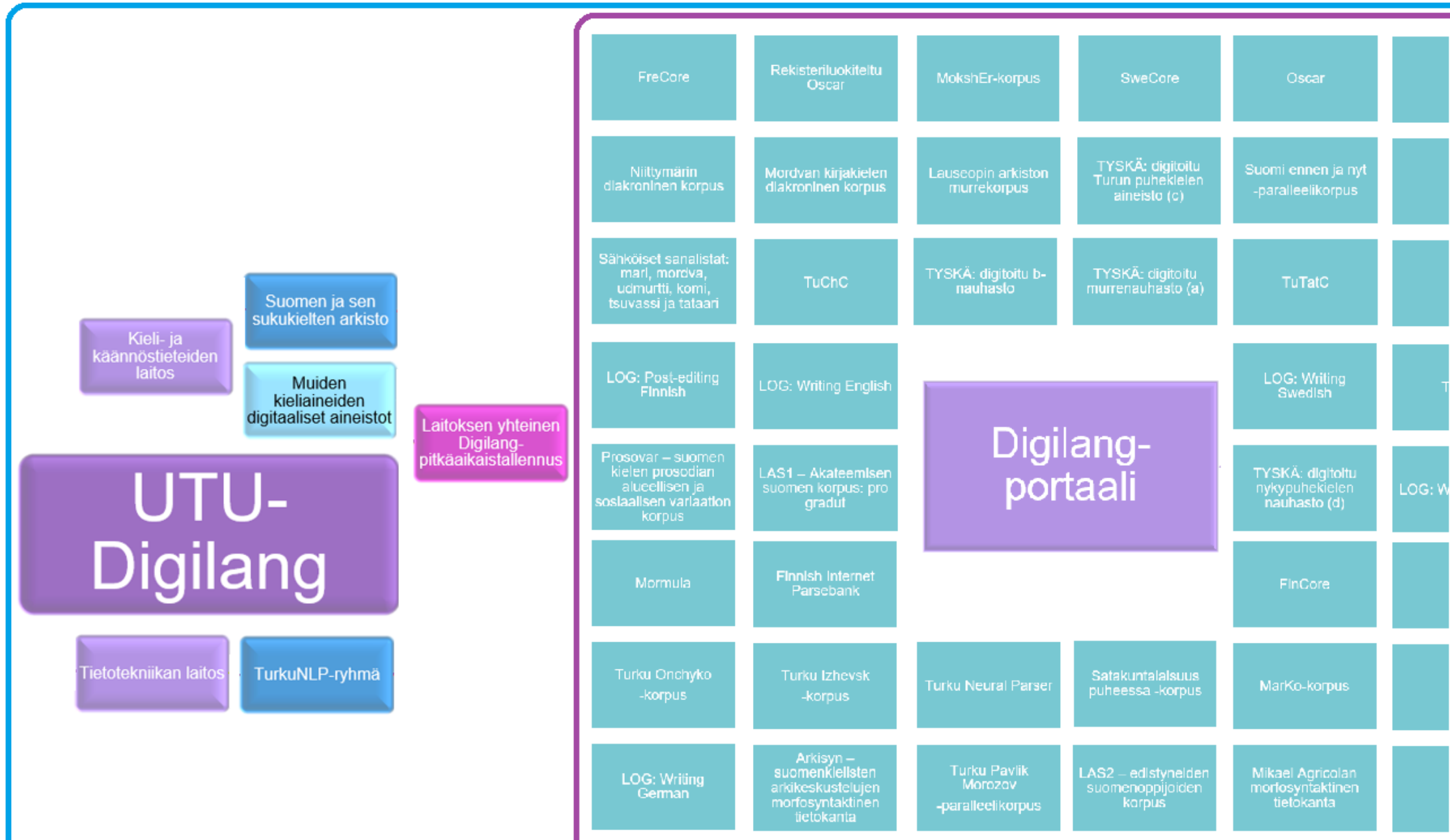


TURUN YLIOPISTO

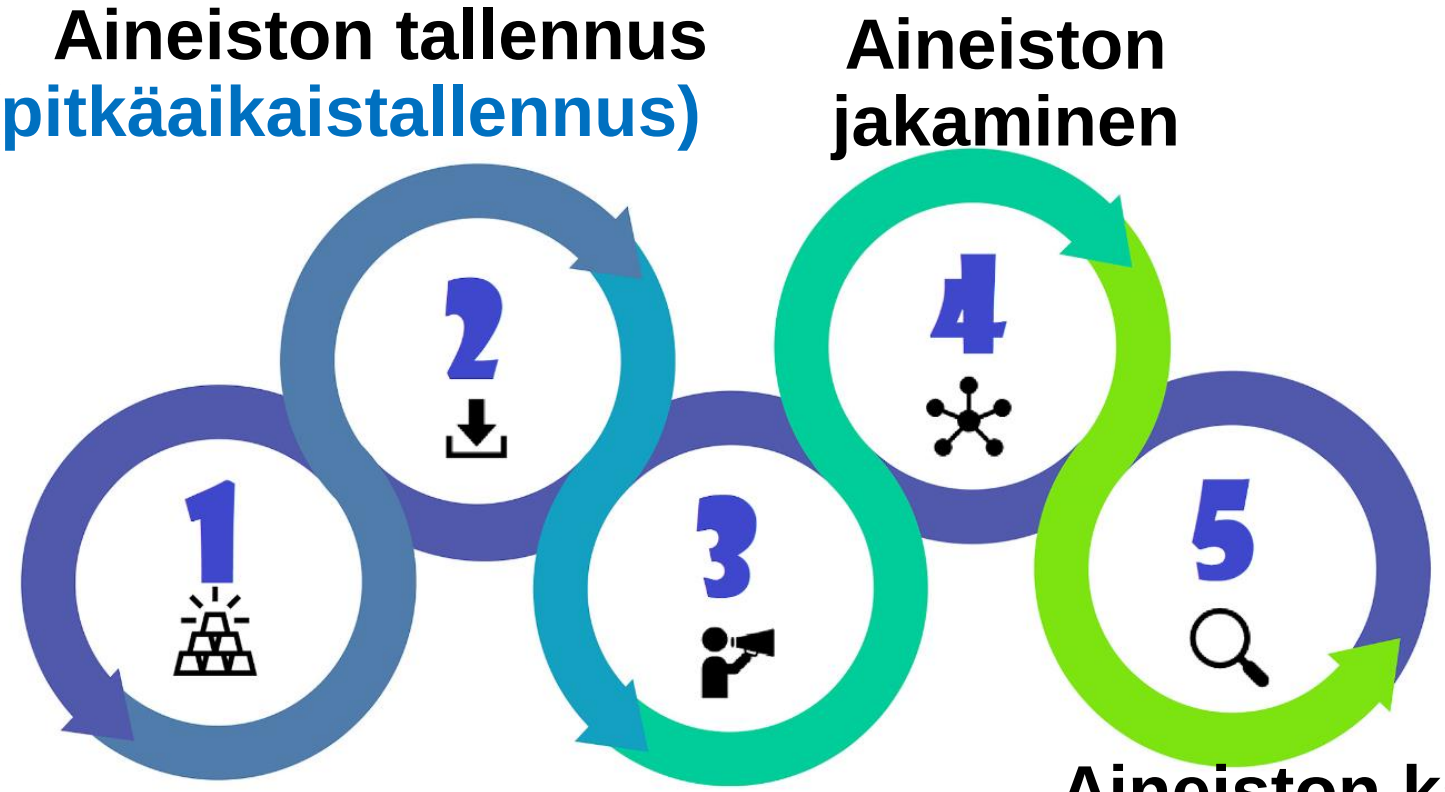
2021



TURUN YLIOPISTO


Kieli- ja käännöstieteiden laitos
Suomen ja sen sukukielten arkisto
Muiden kieliaineiden digitaaliset aineistot
UTU-Digilang
Laitoksen yhteinen Digilang-pitkäaikaistallennus
Tietotekniikan laitos
TurkuNLP-ryhmä
FreCore
Rekisteriluokiteltu Oscar
MokshEr-korpus
SweCore
Oscar
Niittymärin diakroninen korpus
Mordvan kirjakielen diakroninen korpus
Lauseopin arkiston murrekorpus
TYSKÄ: digitoitu Turun puhekielen aineisto (c)
Suomi ennen ja nyt -paralleelikorpus
Sähköiset sanalistat: mari, mordva, udmurtti, komi, tsuvassi ja tataari
TuChC
TYSKÄ: digitoitu b-nauhasto
TYSKÄ: digitoitu murrenauhasto (a)
TuTatC
LOG: Post-editing Finnish
LOG: Writing English
Digilang-portaali
LOG: Writing Swedish
Prosovar – suomen kielen prosodian alueellisen ja sosiaalisen variaation korpus
LAS1 – Akateemisen suomen korpus: pro gradut
TYSKÄ: digitoitu nykypuhekielen nauhasto (d)
Mormula
Finnish Internet Parsebank
FinCore
Turku Onchyko -korpus
Turku Izhevsk -korpus
Turku Neural Parser
Satakuntalaisuus puheessa -korpus
MarKo-korpus
LOG: Writing German
Arkisyn – suomenkielisten arkikeskustelujen morfosyntaktinen tietokanta
Turku Pavlik Morozov -paralleelikorpus
LAS2 – edistyneiden suomenoppijoiden korpus
Mikael Agricolan morfosyntaktinen tietokanta

UTU-Digilang: Digilang-portaali ja Digilang-pitkäaikaistallennus



TURUN YLIOPISTO

UTU-Digilang

https://sites.utu.fi/utu-digilang/

portaali
https://digilang.utu.fi

TURUN YLIOPISTO

UTU-Digilang
Language resource portal

Search by name or keyword

Keywords | Languages | Data Types | Modality | Language Variant Type | L1/L2 | Translated Corpora

Media | Annotations | Text Genre | Dataset Dates | Dataset Size | Availability | Groups

29 datasets | Display list | Show as groups | Arrange

Mormula: Grammatically annotated Mordvin texts (Erzya, Moksha)
Erzya Moksha

Turku 'Pavlik Morozov' Corpus (parallel texts)
Russian Finnish Erzya Moksha Meadow Mari ...

'Finland - Past and Present' Corpus (parallel texts)
Finnish Russian Erzya Moksha Meadow Mari ...

Diachronic Corpus of Literary Mordvin
Erzya Moksha

Diachronic Corpus of Literary Meadow Mari
Meadow Mari

MokshEr Corpus (Moksha, Erzya)
Erzya Moksha

MarKo Corpus (Mari texts)
Mari

Turku Izhevsk Corpus (Udmurt texts)
Udmurt

Turku Onchyko Corpus (Meadow Mari texts)
Mari

Turku Komi-Permyak Corpus (TuKPC)
Komi-Permyak

Turku Tatar Corpus (TuTatC)
Tatar

Turku Chuvash Corpus (TuChC)
Chuvash

Electronic Word Lists: Mari, Mordvin

Mikael Agricolan

Arkisyn –

TURUN YLIOPISTO

UTU-Digilang
Language resource portal

In the same dataset
In the whole databank

- Academic Language
- Academic Writing
- Regional Variation
- Everyday Conversation
- Chuvash
- Dialectology
- Elicited

Keywords | Languages | Data Types | Modality | Language Variant Type | L1/L2 | Translated Corpora

Media | Annotations | Text Genre | Dataset Dates | Dataset Size | Availability | Groups | Archiving Status

29 datasets | Display list | Show as groups | Arrange | Show report

Mormula: Grammatically annotated Mordvin texts (Erzya, Moksha)
Erzya Moksha

Turku 'Pavlik Morozov' Corpus (parallel texts)
Russian Finnish Erzya Moksha Meadow Mari ...

'Finland - Past and Present' Corpus (parallel texts)
Finnish Russian Erzya Moksha Meadow Mari ...

Diachronic Corpus of Literary Mordvin
Erzya Moksha

Diachronic Corpus of Literary Meadow Mari
Meadow Mari

MokshEr Corpus (Moksha, Erzya)
Erzya Moksha

MarKo Corpus (Mari texts)
Mari

Turku Izhevsk Corpus (Udmurt texts)
Udmurt

Turku Onchyko Corpus (Meadow Mari texts)
Mari

Turku Komi-Permyak Corpus (TuKPC)
Komi-Permyak

Turku Tatar Corpus (TuTatC)
Tatar

Turku Chuvash Corpus (TuChC)
Chuvash

Electronic Word Lists:

Mikael Agricolan

Arkisyn –

Diachronic Corpus of Literary Meadow Mari
Meadow Mari

MokshEr Corpus (Moksha, Erzya)

Erzya | Moksha | Literary Language

corpus

Description | Content | Authors | Availability | Referring

Detailed MokshEr Corpus Description

Minimize

MarKo Corpus (Mari texts)
Mari

Turku Izhevsk Corpus (Udmurt texts)
Udmurt

Turku Onchyko Corpus (Meadow Mari texts)
Mari

Turku Komi-Permyak Corpus (TuKPC)
Komi-Permyak

Turku Tatar Corpus (TuTatC)
Tatar

Turku Chuvash Corpus (TuChC)
Chuvash

Electronic Word Lists: Mari, Mordvin, Udmurt, Komi, Chuvash, Tatar
Mari | Moksha | Udmurt | Komi | Chuvash | ...

Satakuntalaisuus puheessa -korpus
Finnish

Mikael Agricolan teosten morfosyntaktinen tietokanta
Finnish

Arkisyn – Suomenkielisten arkikeskustelujen morfosyntaktinen tietokanta
Finnish

LOG: Writing German
German

LOG: Writing French
French

LOG: Writing Finnish
Finnish

LOG: Writing English
English

LOG: Writing Swedish
Swedish

LOG: Post-editing Finnish
Finnish

digilang.utu.fi

Kiitos!

TURUN YLIOPISTO

Suomen ja sen sukukielten arkisto

Tommi Kurki

TURUN YLIOPISTO

Suomen ja sen sukukielten arkisto

- suomen ja sen sukukielten tutkimukseen liittyvän tutkimusaineiston säilyttämiseen ja kehittämiseen erikoistunut arkisto.
- suomen kielen ja suomalais-ugrilaisen kielentutkimuksen oppiaineen yhtenä vahvuusalana aineistoinfrastruktuurien kehittäminen
  - oppiaine onkin kansainvälisesti merkittävin suomen ja sen sukukielten annotoitujen digitaalisten aineistojen tuottaja ja kehittäjä

- Suomen kielen aineistot
- Etäsukukielten aineistot

TURUN YLIOPISTO

Suomen kielen aineistot

- Lauseopin arkisto (LA; 1967-)
  - Puhutun kielen kieliopillisesti annotoidut aineistot
  - Kirjoitetun kielen kieliopillisesti annotoidut aineistot
- Turun yliopiston suomen kielen äänitearkisto (TYSKÄ; 1962-)
  - Ääni- ja videotallenteet sekä niistä niistä kielitieteellisesti transkriboidut litteraatit
  - Em. aineistoista koostetut korpukset
- Kaikki aineistot digitoitu

Etäsukukielten aineistot

- Volgan alueen kielten tutkimusyksikkö
  - Laajoja aineistoja on marista, mordvasta, udmurtista, komipermjakista; tšuvassista ja tataarista.
- Digitaalisten kielikorpusten kokoelma
  - annotoimattomat tekstit (pelkkiä tekstejä ilman kieliopillista analyysia)
  - kieliopillisesti annotoidut tekstit (kaikki sanat analysoitu morfologisesti)
  - paralleelitekstit (sama teksti monella kielellä)
  - kirjakielen historian korpukset (valikoima tekstejä eri vuosikymmeniltä)
  - sanaluettelot (sanat luokiteltuina sanaluokkiin, ei merkityksen selityksiä)

Suomen kielen aineistot

- Kieliopillisesti annotoidut korpukset
  - Lauseopin arkiston murrekorpus
  - Mikael Agricolan morfosyntaktinen tietokanta
  - Arkisyn – suomenkielisten arkikeskustelujen morfosyntaktinen tietokanta
  - Akateemisen suomen korpus (LAS1)
  - Edistyneiden suomenoppijoiden korpus (LAS2)
  - (morfosyntaktinen Satakuntalaisuus puheessa -korpus)
  - (Prosovar – suomen kielen prosodian alueellisen ja sosiaalisen variaation korpus)
- Vuosikymmeniä aineiston jakelua Lauseopin arkistosta, nykyisin jakelu pääasiassa Kielipankin kautta

Etäsukukielten aineistot

- Esimerkkejä etäsukukielten aineistoista
  - Mormula – mordvalaiskielten morfologisesti ja syntaktisesti koodattu korpus
  - Annotoidut tekstikorpukset seitsemästä Volgan alueen kielestä
    - Ersä, moksha, mari, udmurtti, komipermjakki; tshuvassi, tataari
  - Kirjakielten historian korpukset
    - Mari ja mordvalaiskielet
  - Paralleelikorpukset
- Digilang-projektin aikana hankittu aineiston jakelua varten oma palvelin

digilang.utu.fi

Kiitos!

TURUN YLIOPISTO