

Mitä aineistoja ja verkkopalveluja FIN-CLARIN toivoo Kielipankilta?

Aineistojen käsittely- ja käyttöympäristöt (CSC, HY, Aalto, teollisuus)

- tavoitteet (pari riviä):
 - Aineistot: prosessien tehostaminen, jotta pystyisimme laittamaan tarjolle nopeammin kuin tulee uusia aineistoja
 - Palvelut: annotointi- ja hakutyökalujen kehitys, palvelunhallintaprosessien kehitys
 - epästandardin kielen käsittely
- olemassa oleva infra (pari riviä)
 - Kielipankki (Korp, LAT, latauspalvelu, taggerit, META-SHARE, PID:t), laskenta- ja tallennusalueet
- lisäinfra tarpeen (pari riviä):
 - Aineistot:
 - YLE:n karttavat aineistot
 - Suomi24:n karttavat aineistot
 - Internetin karttavat aineistot (määräajoin)
 - Kansalliskirjaston karttavat lehtikokoelmat
 - Kansallisarkiston karttavat digitaaliset kokoelmat
 - Palvelut: työkalujen ja ympäristöjen kehitys:
 - kielen analyysi ja annotointityökalut, esim. NER, Sentiment Analysis
 - mahdollisuus kommentoida hakutuloksia verkossa, esim. WikiMedia
 - annotointityökalu verkossa, esim. WebAnno
 - tallenna hakukonfiguraatio, esim. Korp
 - haun rajausta kartta-alueen mukaan, esim. Korp
 - hakutuloksien tilastollinen analyysi ja visualisointi, esim. Chipster
 - R:n tukeminen ja soveltaminen, esim. Chipster
- kv-yhteistyötahot (pari nimeä): CLARIN, EUDAT, RDA, AARC, PRACE, NeIC
- julkaisu-arvio: (arvio 5-10 vuoden päähän) 150-500
- tutkimusprojektin arvio: (arvio 5-10 vuoden päähän) 15-20
- yvv-vaikutus: (arvio 5-10 vuoden päähän)
 - ihmistieteiden ja historian tutkijoiden bigdata-tutkimus
 - yritysten kuluttajaryhmiä mukailevat palvelut
 - tilastobarometrit

Puhekielen käsittely (Aalto, Kotus, Oulu, HY, Turku, UEF, Kansalliskirjasto)

- tavoitteet (pari riviä):
 - digitoitujen äänitteiden automaattinen tekstitys
 - puhe- ja murreaineistojen käsittely
 - Oulun nauhoitearkiston saaminen FIN-CLARINin piiriin
 - taustatietojen puoliautomaattinen lisääminen äänitteisiin
 - Kotuksen kannalta tärkeintä olisi, jos syntyisi tekniikkoja automaattisesti käsitellä murrepuhetta, esim. koneellinen raakalitterointi yms.
 - Tavoitteena voisi olla, että **kaikki digitaaliset av-aineistot saadaan Kielipankin kautta tutkijoiden käyttöön**
 - aineistojen **annotointia** edistetään eri menetelmin.
 - arkikielen korpus ja sen litterointi (500 puhujaa x 5 min)
- olemassa oleva infra (pari riviä)
 - Aineistot:
 - Prosodiaprojektin ja SKN:n näytteitä koko Suomesta
 - Arkisynin ja lauseopin murrekorpuksen puhekielistä keskustelua
 - kerätty aineistoja mitä on saatavilla
- lisäinfra tarpeen (pari riviä):
 - Palvelu: adaptiivinen puheentunnistuspalvelu (ASR)
 - Joidenkin aineistojen luvittaminen jälkikäteen (GDPR)
- kv-yhteistyötahot (pari nimeä): ICSI (USA), NINJAL (National Institute for Japanese Language and Linguistics) (balanssoitu korpus keskustelutilanteista: surveyn perusteella), MeMAD (esim. EURECOM, INA), Edinburgh, Saarland
- julkaisu-arvio: (arvio 5-10 vuoden päähän) 10-20
- tutkimusprojekti-arvio: (arvio 5-10 vuoden päähän) 5-10
- yvv-vaikutus: (arvio 5-10 vuoden päähän)
 - FCAI (Aalto ja HY on mukana)
 - Intersubjektiivisuus vuorovaikutuksessa -tutkimushanke Suomen Akatemian tutkimuksen huippuyksiköksi vuosiksi 2012–2017. Hankkeessa tutkitaan, miten keskustelu ja muu sosiaalinen vuorovaikutus synnyttää ihmisten välisen ymmärtämisen ja mahdollistaa kokemusten jakamisen.

Videoiden ja kuvien käsittely (Aalto, Kotus, Kuurojen liitto, Kansalliskirjasto, Kansallisarkisto, HY)

- tavoitteet (pari riviä):
 - Palvelut:
 - kuvailutulkkaukset ja -tekstitys kuvien ja videoiden haettavuuden edistämiseksi (EU-hanke MeMAD: Aalto, HY)
 - puheen- tai viittomien automaattinen raakatunnistus ja -annotointi
 - puoliautomaattinen työkalu kuvailutulkkauksen tuottamiseen (2020 loppuun mennessä)
- olemassa oleva infra (pari riviä)
 - Aineistot:
 - Kuurojen liiton annotoidut aineistot

- Aallon videoaineistot (pienimuotoisia)
 - TaY:ssa kootaan tulkkaustilauksuuksien videonauhoituksia, tulkkeina opiskelijoita. Myös tulkkaus puhelimitse. Kielipari suomi-venäjä. Aineistot litteroidaan ELAN:illa, tällä hetkellä 12 nauhoitusta X noin 30 min.
 - YLEn aineistot suomi-ruotsi-tekstitys
 - Kikosa-aineisto (20 h monikielinen video ja puhe)
 - MUTABLE
- lisäinfra tarpeen (pari riviä):
 - Aineistot:
 - FIN-CLARIN on jakelukanava tutkimuskäyttöön yhteistyössä YLEn ja KAVIn kanssa
 - FIN-CLARINilta edellytykset jakaa videoaineistoa tutkijoille: API-liitännän rakentamista ja "Kopioisto"-sopimuksia
 - tutkijoiden omat aineistot samaan käsittelyyn
 - YLEn tekstitetyt videoaineistot
 - annotoidut kuvailutulkkausaineistot kehitys- ja testaustarkoituksiin
 - Audio ja videosisällön automaattinen ja puoliautomaattinen kuvailu
 - Tarvitaan tuki LATissa kaksikielisille aineistoille
- kv-yhteistyötahot (pari nimeä): Univ. Surrey (uk), New York Univ. (usa),
- julkaisuarvio:(arvio 5-10 vuoden päähän):25-50
- tutkimusprojektiarvio: (arvio 5-10 vuoden päähän): 10-15
- yvv-vaikutus: (arvio 5-10 vuoden päähän):
 - nyt voitaisiin tehdä kvantitatiivista tutkimusta, kun aikaisemmin vain kvalitatiivista mediatutkijoiden piirissä
 - FCAI

Fraktuura- ja käsin kirjoitettujen dokumenttien käsittely (Kotus, Kansallisarkisto, Kansalliskirjasto)

- tavoitteet (pari riviä):
 - kirje- ja nimiarkistojen sähköinen saavutettavuus parannetaan tarjoamalla palvelu joka paitsi skannaa myös tunnistaa mitä käsin on kirjoitettu
 - sopivaa materiaaliahan Kotuksessa jo on digitaalisena (mm. *Nimiarkisto*) ja kehittynyt tekstintunnistus loisi aivan uudenlaisia mahdollisuuksia myös vaikkapa *murrearkiston* digitointiin
 - *paperiarkistot* digitalisoidaan ja että tässä hyödynnetään nimenomaan tekstintunnistusta ja osallistutaan tunnistusmenetelmien kehittämiseen.
- olemassa oleva infra (pari riviä)
 - Aineistot: Kotus: nimiarkisto, KA: kirjekokoelmat, KK: kuva- ja lehtikokoelmat
 - Kansallisarkisto kehittää EU-tasolla jo tällaista palvelua READ-projektin puitteissa
 - FIN-CLARIN voi tukea tätä Suomessa saadakseen lisääaineistoja ja tarjotakseen vastaavaa palvelua jäsenorganisaatioilleen

- lisäinfra tarpeen (pari riviä):
 - Palvelut:
 - joukkoistusliittymä
 - käsialan tunnistuspalvelu (READ)
 - vanhojen fraktuura-aineistojen jälkikäsitteily (OCR ja post-OCR)
 - käsialan- ja kuvantunnistusta (AHR)
- kv-yhteistyötahot (pari nimeä):
 - READ-projekti (TU Wien), TranScriptorium (UCL)
- julkaisu-arvio: (arvio 5-10 vuoden päähän) 100-500 (eri puolella Suomea)
- tutkimusprojekti-arvio: (arvio 5-10 vuoden päähän) 10-20 (ympäri Suomea)
- yvv-vaikutus: (arvio 5-10 vuoden päähän)
 - kotimaisten skannauspalvelujen kehittyminen
 - historian tuntemus kouluissa
 - suku- ja kotiseutututkijoiden historiikit

Oppijoiden korpuksia (Oulu, Jyväskylä, Turku, Kotus, Aalto)

- tavoitteet (pari riviä):
 - oppijoiden puhe- ja kirjoitussuoritusten automaattinen annotointi (hyöty opetukselle ja arvioinnille);
 - taitotason mahdollinen päättely annotoidun aineiston avulla;
 - virheannotoinnin täydentäminen (automaattisen virheannotoinnin kehittäminen)
- olemassa oleva infra (pari riviä)
 - Aineistot: LAS, ICLFI - kielen oppijakorpuksia
 - CEFLING, TOPLING ja DIALUKI -oppijakorpuksia
 - nykyinen väline: täysin käsinannotoitu virheannotointi
 - morfosyntaksi puoliautomaattinen olemassa
 - abien puhumaa ruotsia ja lasten puhumaa englantia (Aalto)
 - Akateeminen suomi kv-oppijan haasteena (TuY strategiarahoitus) osa Digilang-hanketta
 -
 - [HUOM! Digilang sisältää myös Prosodiakorpus, Arkisyn, ParseBank, FennoUgri-aineistot.]
- lisäinfra tarpeen (pari riviä):
 - Palvelut:
 - kielellisen annotoinnin verkkotyökalu; virheannotoinnin verkkotyökalu
 - työkalu tekstien ominaisuuksien ja vaikeustason arviointiin, esim. coh-metrix (englannin kielessä)
 - lisämahdollisuuksia hyödyntää annotointia haun yhteydessä
 - myös muunkieliset aineistot kuin suomi tarpeen (engl., ruots.)
 - suullisen taidon arviointivälineet ja arviointituki
 - Suullisen taidon ja ääntämisen automaattinen arviointi puheentunnistuksen avulla eri kielille

- kv-yhteistyötahot (pari nimeä): Lancaster University, Penn State University, Reitaku University (Japan), TalTech, Tallinn University
- julkaisu-arvio: (arvio 5-10 vuoden päähän): 40-80
- tutkimusprojekti-arvio: (arvio 5-10 vuoden päähän): 5-8
- yvv-vaikutus: (arvio 5-10 vuoden päähän):
 - oppijakielen analyysivälineiden tuki opetukselle ja oppimista tukevalle formatiiviselle arvioinnille; datapohjaisen opettamisen ja oppimisen kehittäminen; oppimateriaalityö
 - Kielenopettajien harjoitusympäristö suullisen taidon arviointiin
 - Kielenopiskelijoiden harjoitusympäristö suulliseen kielitaitoon

Käännösaineistot (Tampere, UEF, HY)

- tavoitteet (pari riviä):
 - Palvelu: käännösaineistojen haku ja käyttö
 - Kollokaatioiden haku, frekvenssilistat, vaikea ladata tulokset, ...
 - monikieliset työkalut ja luvallisten aineistojen haravointi
- olemassa oleva infra (pari riviä)
 - Aineistot:
 - ParRus ja ParFin, Opus ja Kirjallisuuspankki
 - kaksikieliset aineistot
 - yhteistyö HY:n kanssa (kaunokirjallisuus) [Kopotev, Nikunlassi (HY)]
 - suomen ja venäjän/ruotsin välille tehdään rinnakkaiskorpusta (lakitekstiä: Valitiosopimusten rinnakkais tekstikorpusta. Tällä hetkellä koottu ja kohdistettu noin 120 tekstiparia suomi-venäjä.)
 - LAT-käyttöliittymä voi olla hyödyksi (monikielinen tuki tarpeen)
- lisäinfra tarpeen (pari riviä):
 - Korpin kehitystä paremmin käyttäjiä palvelevaksi
 - Hakutulosten lataus omalle koneelle (esim. csv-formaatissa)
 - Isompi ikkuna tulosten näyttämiseksi
 - Rinnakkaiskontekstit eri sarakkeissa
 - Frekvenssilistat
 - Kollokaatioiden haku
 - asioimistulkkauksia (esim. lääkärin, maahanmuuttajakonteksti): kielentunnistus
 - työelämäntulkkauksia (esim. monikieliset organisaatiot): vrt. puheaineiston infra
- kv-yhteistyötahot (pari nimeä): EMT-verkosto EU-rahoitus (Horizon 2020), Venäjän kansalliskorpus (Russkij natsionalnyj korpus, ruscorpora.ru).
- julkaisu-arvio: (arvio 5-10 vuoden päähän): 30-50
- tutkimusprojekti-arvio: (arvio 5-10 vuoden päähän): 10-30
- yvv-vaikutus: (arvio 5-10 vuoden päähän):
 - maahanmuuttajien integrointi
 - tulkkauksetutkimuksen ja käännöstutkimuksen teknologinen suuntaus sekä kääntäjien ja tulkkien koulutus tukee alan työelämävalmiuksia

Sanakirjojen kehitystä ja välittämistä (Kotus, UEF, Turku, Tampere, HY)

- tavoitteet (pari riviä):
 - Lisää sanakirjoja verkkoon ja joukkoistamiseen
 - Sanakirjat ja sanahaku (yhdessä Koneensäätiön ja Suomalaisugrilaisen seuran kanssa)
 - lakitermejä Suomi-Venäjä (ELY-keskuksen kautta rahoitusta, TaY, yhteistyötä HYN ja JoYN kanssa)
 - **sanat.csc.fi**-palveluun sukukielten sanastoja, taajuuslistoja ym.,
 - kieliteknologialla olisi paljon annettavaa, jos tulevaisuudessa luodaan sanakirjoja ja aineistoja yhdistävää kokonaisuutta (ns. AISA).
 - **Kotuksen sähköiset aineistot ja sähköisten sanakirjojen julkaiseminen kokonaisuudessaan**
- olemassa oleva infra (pari riviä)
 - Aineistot ja sanakirjat: fenno-ugristen kielten (Turku) ja Karjalan kielen (UEF)
 - Suomalais-ugrilaiset kielet ja internet (Koneen säätiö) 90 M sanaa
- lisäinfra tarpeen (pari riviä):
 - Palvelu: sanat.csc.fi
 - Volgan alueen kielten tutkimusyksikkö (aineistot siirrettävä Kielipankkiin)
- kv-yhteistyötahot (pari nimeä): Venäjän tiedeakatemia, **Tromssa, Hamburg**, sekä muut maat missä fenno-ugristiikka, ELEXIS (Kotus tarkkailijana)
- julkaisuarvio (arvio 5-10 vuoden päähän): 50-200
- tutkimusprojektiarvio (arvio 5-10 vuoden päähän): 5-10
- yvv-vaikutus (arvio 5-10 vuoden päähän):
 - uhanalaisten kielten tutkimus voi tukea kielten puhujien asemaa yhteiskunnassa

Termien tunnistus ja erikoiskielen käyttö (Vaasa, HY, Turku, teollisuus)

- tavoitteet (pari riviä):
 - Palvelu:
 - terminologian näyttäminen kontekstissa
 - tukea Tieteen termipankkia ja termitutkijoita kieliteknologisesta näkökulmasta
 - tietokoneohjelmien ja tietojärjestelmien käyttöliittymien termien ja muiden kielellisten elementtien tutkimista varten korpuksen luominen
 - käsitesuhdetiedon poiminta teksteistä käsitejärjestelmien, ontologioiden ja sanastojen rakentamiseksi eri aloilta
 - suomen kieleen liittyvien terminologisten resurssien löydettävyyden ja saatavuuden tuki ja käyttö, palvelun ja hakutulosten seuranta sekä niitten pohjalta palvelun kehittäminen
- olemassa oleva infra (pari riviä)

- Aineistot: tiedelehdet, verkossa olevia termipankkeja
- Käsin koottua termistöä muutaman yhteistyöyrityksen ohjelmiston käyttöliittymistä (Wilma, Jakamo)
- käsitesuhteisiin liittyvä taksonomia
- Laadukkaita sanastoja/termistöjä Kielipankkiin (muilla aloilla kuin Tieteen termipankin edustamilla aloilla)
- lisäinfra tarpeen (pari riviä):
 - Aineistot: opinnäytteet
 - Palvelu: termiesimerkkien poiminta Korp API:n kautta
 - Kansainväliset lisäinfraesimerkit: kartoitus (Italia, Portugali, Norja, Tanska)
 - Aineisto: tietokoneohjelmien ja tietojärjestelmien käyttöliittymien kielellisestä aineistosta korpus, palvelu: terminpoiminta käyttöliittymistä, tekijänoikeusapua
 - työkalu käsitesuhdetiedon poiminnan automatisoimiseksi
- kv-yhteistyötahot (pari nimeä): Eurac Research Italia ([Institute for Specialised Communication and Multilingualism](#)) tiedonvisualisoinnin termityökaluja; NHH Bergen Norja; CBS Kööpenhamina Tanska, EU TermCoord, EAFT-järjestö
- julkaisuarvio:(arvio 5-10 vuoden päähän) 20-30
- tutkimusprojektiarvio: (arvio 5-10 vuoden päähän) 4-5
- yvv-vaikutus: (arvio 5-10 vuoden päähän)
 - erikoisalojen sanastojen ja sanastotyön kehittäminen; suomen kielisen tieteentekemisen tukeminen; termipankkiosaamisen levittäminen yrityksiin, tietokoneohjelmien ja tietojärjestelmien kielellisen ulkoasun ohjeistuksen luominen ja käyttöliittymien suunnittelun ja lokalisoinnin laadun nostaminen