þÿ K i e l i p a n k k i a i n e i s t o j e n j a t y ö k a l u j e n r u n s a u d e

Lennes, Mietta

2021-12

þÿ L e n n e s , M 2 0 2 1 , ' K i e l i p a n k k i a i n e i s t o j e n j a t y ö k a l u j e n r u n s a u d e n s a
ruotsinsuomalainen kielenhuollon tiedotuslehti , Nro 4/2021 , Sivut 4-9 .

http://hdl.handle.net/10138/356999

unspecified
publishedVersion

Kieliviesti

Suomi ja meänkieli Ruotsissa 4 • 2021

Kielipankki

Ehdotukset vähemmistökielten vahvistamiseksi

Kieliraatin meänkieliseminaari 2021

Kieliviesti

Suomi ja meänkieli Ruotsissa 4 • 2021

Ilmestyy neljästi vuodessa.
Tilaushinta vuonna 2021 Ruotsissa: 200 kr (sis. alv:n)
Tilaushinta vuonna 2021 ulkomaille: 280 kr (sis. alv:n)
Vuodesta 2022 alkaen lehti muuttuu maksuttomaksi digilehdeksi.

Päätoimittaja	Riina Heikkilä
Toimituskunta	Tarja Larsson
	Sari Pesonen
	Elina Kangas (meänkieli)
Ulkoasu	Anna Jemsö
Postiosoite	Kielineuvosto, Box 20057, 104 60 Stockholm
Katuosoite	Alsnögatan 7
Puhelin	0200-275 555 (klo 9–12)
Sähköposti	suomi@isof.se
	(myös tilaukset ja osoitteenmuutokset)
Verkkosivusto	www.isof.se

ISSN 0280-350X

Kielineuvosto
Kielen ja kansanperinteen tutkimuslaitos

Institutet för språk och folkminnen

Paino: Stibo Complete

Mietta Lennes

Kielipankki – aineistojen ja työkalujen runsaudensarvi

Vuonna 1996 perustettu Kielipankki on erityisesti tutkijoille suunnattu palvelukokonaisuus, jossa on tarjolla laaja valikoima erikielisiä teksti- ja puheaineistoja. Nimestään huolimatta Kielipankin tarjontaa voivat hyödyntää muutkin kuin kielitieteilijät. Suuri osa Kielipankin kautta välitettävistä aineistoista on julkisesti kenen tahansa saatavilla ja niistä voi tehdä monipuolisia hakuja verkkoselaimella. Useimmat aineistot voi tarvittaessa myös ladata omalle koneelle tutkittaviksi.

Kielipankin verkkosivuilla olevasta aineistoluettelosta *(www.kielipankki.fi/aineistot)* näkyy, millaisilla ehdoilla ja missä palvelussa kutakin aineistoa pääsee käyttämään. Kielipankista löytyy myös työkaluja *(www.kielipankki.fi/tyokalut)*, joilla käyttäjät voivat analysoida ja käsitellä omia aineistojaan.

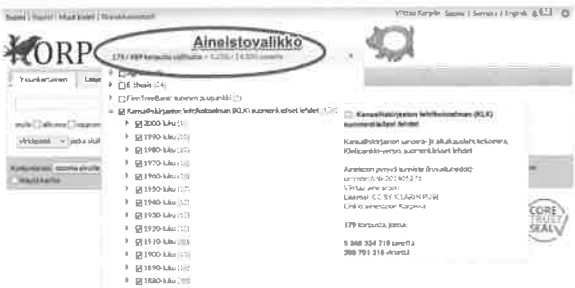
Etenkin ihmistieteellisessä tutkimuksessa tarvitaan nykyisin yhä useammin valtavia digitaalisia aineistoja, jotka sisältävät tekstiä tai puhetta. Tutkimusta varten tarvittavan teksti- tai puheaineiston kerääminen, järjestäminen, esikäsittely ja dokumentointi on usein työlästä ja sama aineisto saattaa sopia moniin erilaisiin tutkimusaiheisiin. Siksi on hyvä, että tutkijat voivat sijoittaa aineistonsa Kielipankin tapaiseen paikkaan, josta niitä on mahdollista välittää eteenpäin.

Esimerkiksi suomalaiseen Kansalliskirjastoon tallennetut digitaaliset sanoma- ja aikakauslehtikokoelmat tarjoavat hienoja

Piirros: Markku Huovila

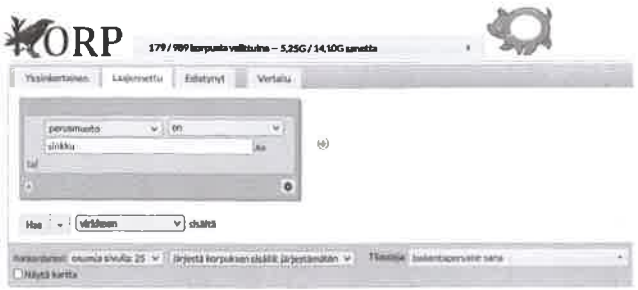
mahdollisuuksia historian ja yhteiskuntatieteiden tutkijoille. Laajojen tekstiaineistojen käsittely ja tutkiminen olisi kuitenkin todella työlästä ja hankalaa, jos yksittäisten artikkeleiden sisällöstä ei voisi saada mitään käsitystä lukematta jokaista dokumenttia ihmisvoimin lävitse. Onkin kiinnostavaa huomata, kuinka paljon aineistojen käyttämistä ja tutkimista voidaan helpottaa automaattisilla menetelmillä.

Jos esimerkiksi haluaisi etsiä tavallisesta suomenkielisestä tekstistä kaikki kohdat, joissa mainitaan sana käsi, pitäisi luultavasti hakea yksitellen eri muotoja *käsi, käden, kättä* jne. Kun teksti on etukäteen jäsennetty, sanan kaikkien muotojen esiintymät voidaan kerätä kohdistamalla haku sellaisiin



Kuva 1: Kansalliskirjaston lehtikokoelmasta tuotetun KLK-aineiston valitseminen käyttöön Kielipankin Korp-palvelussa

sanoihin, joiden perusmuodoksi on merkitty *'käsi'*. Jäsennetyt aineistot ja hakua helpottavat välineet ovat erityisen tärkeitä suomen kaltaisille kielille, joissa sanoja voidaan tai-



Kuva 2: Esimerkkihaku, jolla etsitään sanan 'sinkku' esiintymiä missä tahansa muodossa. Tässä haun kohteeksi oli valittu KLK-aineisto (Kansalliskirjasto, 2011).

KLK SUOMI 2000 (ei tue laajennettua kontekstia)

omantt., naisell., 3-kympp., vapaa,	**sinkku**	nainen pääkaupunkiseud. kaipallee vastaavanlaista n
nttyyt yhteen Torremolinos 2000	**-sinkun**	ja kiertueen merkeissä.
iccinoa lutkuttavien epävarmojen	**sinkkujen**	kollektiivilla on kuitenkin muutama seikka puolellaa
Kuopiolainen nainen n. 30-40- v.	**sinkku**	/ yh.
Elää	**sinkkuna**	L Assa.
Danny kauppaa	**sinkuille**	ohjelmoitavia hakulaitteita.
Siitä tuli Vanha holvikirkko -	**sinkun**	käsintöpuoli.
	sinkkuja	koskeva tilastollinen havainto on aivan tuore.
hmettelee, miksi yksinasuvien ns.	**sinkkujen**	vastaukset ovat negatiivisempia kuin lapsiperheiden
hoitanut bändi pääsee tekemään	**sinkkua**	Xonin kustannuksella.
arrastusseuraa etsii keskikäinen	**sinkku**	nainen. ¥75272
Tänä vuonna voisi ainakin	**sinkun**	tehdä, Aija kaavallee.
Bändin	**Sinkku**	JOIN ME ja albumi RAZORBLADE ROMANCE ovat t
unohdettu	**sinkkujakaan**,	onhan tapahtuman ykkösjärjestäjäkin sinkku.
n tapahtuman ykkösjärjestäjäkin	**sinkku**	.
Kaksi	**sinkkua**	tapasi pikkikisoissa, ja ensi kesänä he menevät naimi

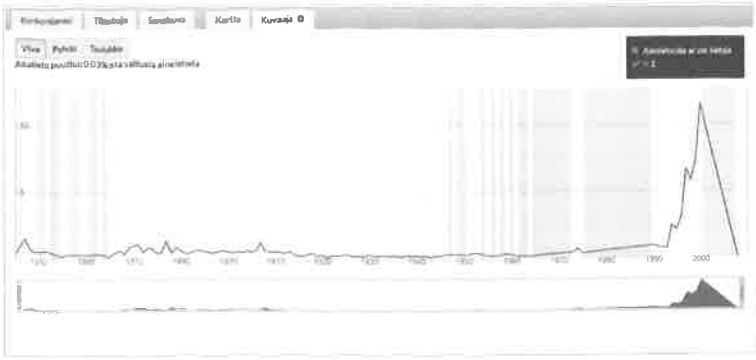
Kuva 3: Sinkku-sanan esiintymiä KLK-aineistoon sisältyvissä lehdissä vuodelta 2000.

vuttaa monin tavoin. Kehittynyt kieliteknologia voi siis auttaa aineiston hallinnassa ja hakujen tekemisessä, vaikka käyttäjä ei itse olisikaan sinänsä kiinnostunut esimerkiksi sanojen rakenteesta tai lauseiden kieliopillisista piirteistä.

Kokeile Korpia

Helpoiten Kielipankin aineistoja pääsee kokeilemaan ja tutkimaan Korp-nimisessä hakupalvelussa. Korp-alustaa kehitetään Göteborgin yliopiston Språkbankenissa, mutta myös suomalaisessa Kielipankissa voi käyttää samaa työkalua. Eri maiden kielipankit tekevät siis keskenään yhteistyötä. Kielipankin Korp-alusta sijaitsee osoitteessa *https://korp.csc.fi*. Språkbanken tarjoaa Korp-alustallaan pääsyn moniin laajoihin ruotsinkielisiin aineistoihin, kun taas suomalaisen Kielipankin ylläpitämästä Korpista löytyy enemmän suomen ja suomenruotsin aineistoja. Korp-palvelun käyttöohjeita ja hakuesimerkkejä löytyy suomeksi Kielipankin verkkosivuilta osoitteesta *https://www.kielipankki.fi/tuki/korp*.

Aineistojen tutkiminen on kiinnostavaa salapoliisityötä. Yksittäisen sanan erilaisia



Kuva 4: Sinkku-sanan esiintymistä laskettu trendikuvaaja KLK-aineistossa (Kansalliskirjasto, 2011).

	Sinkkuja	otetaan sauhutettavaksi kuuman ja kylmän sauhun kanssa.
	Sinkkuja	ja kaikenlaisia ruokatavaroita kohtuhintoihin.
Kallio 3 linja 19. Hyvää maalaisvoita.	**sinkkuja**	, lihaa, munia Ja omenia y. m halvimmilla hinnoilla.
ıuntaina matkalla Tukholmasta Hel »	**sinkun**	sattui lähellä Tammisaarta klo puoli 2 päiwällä onnettomuus
Janstoa, Saksan Sipolla, Sauhutettua	**Sinkkua**	, Heliumin Makkaraa.
› kusti samana iltana Pietarista Hel »	**sinkun**	, jossa sitten eduskunta seu< raawana päiwänä eli maanantai
Hil »	**sinkun**	.
s myöskin kenraali siksi saapuu Hel »	**sinkun**	.
:^iä, että nwosiwoitosta 444,891: 29	**Sinkusta**	on ensin! mäheii' nettäwä 10 « o 3mf.
ar » tauähettH tlo puoli 3 i. p. paätJti	**Sinkku**	jtt lioi 6 f. p. Sainion maimaistalolla liri » töherrä Hannula.

Kuva 5: Sinkku-sanan esiintymiä KLK-aineistoon sisältyvissä lehdissä vuodelta 1909.

merkityksiä ei välttämättä tule edes ajatelleeksi, ennen kuin näkee, millaisissa yhteyksissä niitä on todellisuudessa käytetty. Saman näköistä sanaa voidaan käyttää eri merkityksissä eri aikoina. Kun tutkitaan vaikkapa sanan 'sinkku' esiintymiä Kansalliskirjaston sanoma- ja aikakauslehtiaineistoissa (ks. kuvat 1-2), löytyy 2000-luvun lehdistä esimerkkejä, joissa kyseinen sana viittaa ilman parisuhdetta elävään henkilöön (kuva 3). Musiikkiin liittyvissä käyttöyhteyksissä 'sinkku' voi toisaalta tarkoittaa single-levyä.

Hieman yllättäen vaikuttaa Korpin tarjoamien tilastotietojen mukaan siltä, että 'sinkku'-sana olisi esiintynyt suomenkielisissä lehdissä jo 1800-luvulla ja 1900-luvun alkupuolella (kuva 4). Kun tuolta ajalta tutkitaan yksittäisiä esiintymiä lähemmin, huomataan, että osa "sinkun"-muodoista on ilmeisesti automaattisessa kuvantunnistuksessa tapahtuneita virheitä, joissa alkuperäisen sanomalehden sivulla ollut sana "Helsinkiin" on muuttunut muotoon "Hel" ja "sinkun" (kuva 5). Virheiden lomasta löytyy kuitenkin myös todisteita siitä, että esimerkiksi vuonna 1909 *sinkku*-sanalla onkin tarkoitettu *kinkkua*.

Korpin avulla voi haluamistaan tekstiaineistoista etsiä yksittäisiä sanoja tai useampisanaisia ilmauksia ja rakenteita sekä tarkastella niiden käyttöyhteyksiä tai tilastoida ja vertailla niiden esiintymistaajuuksia. Korpissa näkyvät aineistot on Kielipankissa esikäsitelty käyttäjiä varten. Kielitieteellisen jäsennyksen ohella aineistoihin voidaan joko käsin tai koneellisesti lisätä myös muita merkkauksia, jotka helpottavat ja nopeuttavat sisällön tutkimista. Esimerkiksi tekstissä esiintyvät nimet on saatettu automaattisesti tunnistaa, aineistoon on saatettu lisätä paikkatietoja tms. Yksittäiseen korpukseen sisältyviin dokumentteihin voi myös liittyä hyödyllisiä kuvailutietoja, esimerkiksi tieto kunkin tekstin julkaisuajankohdasta, jota voi käyttää hyväksi hakujen rajaamisessa.

Aineistoja käyttäessä on muistettava, ettei automaattinen jäsennys ole virheetöntä. Tehokkaiden työkalujen avulla hakuja ja analyyseja on onneksi mahdollista toistaa, tarkistaa ja parannella eri näkökulmista. Toistettavuus onkin tieteellisessä tutkimuksessa erityisen tärkeää. Kielipankki tarjoaa jokaiselle aineistoversiolle kuvailutiedot sekä yksilöllisen tunnisteen ja viittausohjeen tutkijoita varten.

Aineistoja moniin tarpeisiin

Kielipankin valikoimissa on eri kokoisia ja eri tarkoituksiin koostettuja aineistoja. Tavoitteena on tuoda aineistot saataville niin avoimesti kuin mahdollista. Osa sisällöistä voidaan kuitenkin tarjota vain yliopiston myöntämällä tunnuksella kirjautuneille käyttäjille, ja pääsy erikseen suojattuihin aineistoihin annetaan vain hakemuksesta. Pääsyrajoituksia ja suojatoimia saatetaan tarvita esimerkiksi sellaisille teksti- ja puheaineistoille, jotka sisältävät henkilötietoja tai tekijänoikeuksien alaista materiaalia. Kunkin aineiston välitysehdoista sovitaan erikseen tutkijan, tutkimusryhmän tai muun tahon kanssa.

Osa Kielipankin aineistoista on tutkijoiden tai tutkimusryhmien keräämiä. Kielipankki on lisäksi keskitetysti järjestänyt tutkimuskäyttöön sellaisia suuria aineistoja, joiden kerääminen ja hallinnointi olisi yksittäisille tutkimusryhmille hankalaa. Tällaisia ovat esimerkiksi *Suomi 24* -aineisto ja *Kansalliskirjaston sanoma- ja aikakauslehtikokoelma*. Eduskunnan täysistunnot pohjautuu eduskunnan julkaisemiin täysistuntopöytäkirjoihin ja täysistuntojen videoihin. Pöytäkirjojen tekstit on Aalto-yliopistossa kohdistettu videoihin automaattisella puheentunnistusmenetelmällä. Korp-palvelussa voi tehdä hakuja näin syntyneestä tekstistä, tilastoida tuloksia eri tavoin sekä katsoa videoita hakutulosten kohdalta.

Edellä mainittujen aineistojen lisäksi Kielipankissa on tarjolla esimerkiksi suo-

men murteiden aineistoja, viittomakielisiä aineistoja, käännöskorpuksia ja monenlaisia sanastoja. Suomenkielisten korpusten ohella aineistoja löytyy ruotsiksi, englanniksi ja monilla muilla kielillä.

Kielipankki ei toimisi ilman yhteistyötä

Kielipankkia koordinoi Helsingin yliopisto, mutta palveluiden ja aineistojen valikoimaa rakentaa yhteisesti suomalainen FIN-CLARIN-konsortio, jonka jäseniä ovat Helsingin, Itä-Suomen, Jyväskylän, Oulun, Tampereen, Turun ja Vaasan yliopistot, Aalto-yliopisto sekä Kotimaisten kielten keskus ja CSC – Tieteen tietotekniikan keskus. FIN-CLARIN puolestaan edustaa Suomea yhteiseurooppalaisessa CLARIN ERIC -nimisessä tutkimusinfrastruktuurissa, jonka kautta suomalaiset tutkijat pääsevät käyttämään muissa maissa koostettuja tutkimusaineistoja ja kielipankkeja.

Kielipankin Kuukauden tutkija -haastattelusarjassa esitellään eri alojen tutkijoita, jotka ovat hyödyntäneet Kielipankissa olevia aineistoja. Vuodesta 2016 alkaen kerättyjä tutkijoiden haastatteluja voi selata osoitteessa *https://www.kielipankki.fi/kielipankki/kuukauden-tutkija-arkisto/.*

Lahjoita puhetta

Lahjoita puhetta on Ylen, Valtion kehitysyhtiö Vaken (nykyinen Ilmastorahasto), Solita Oy:n ja Helsingin yliopiston yhteinen hanke, jonka tavoitteena on kerätä arkista suomenkielistä puhetta kielentutkimuksen ja tekoälyn kehittämisen tarpeisiin. Hankkeen tukena on ollut myös kielen ja puheteknologian asiantuntijoita muun muassa Aalto-yliopistosta ja Turun yliopistosta. Kesäkuussa 2020 alkaneen kampanjan aikana suomenkielistä puhetta on lahjoitettu jo noin 4000 tuntia. Kerätty aineisto tallennetaan Kielipankkiin, jonka kautta lahjoitettua puhetta voidaan tietyillä ehdoilla luovuttaa tutkijoille ja yrityksille.

Jos haluat, voit itsekin lahjoittaa puhettasi nimettömänä esimerkiksi puhelimella tai tietokoneella. Lahjoituksia voi tehdä suomeksi osoitteessa https://lahjoitapuhetta.fi/. Marraskuussa on käynnistynyt myös suomenruotsin kampanja, jossa puhettaan voi lahjoittaa myös ruotsiksi osoitteessa https://doneraprat.fi/. ✍

Kirjoittaja toimii projektisuunnittelijana Helsingin yliopiston digitaalisten ihmsitieteiden osastolla.

Aineistolähde

"KLK-aineisto": Kansalliskirjasto (2011). *Kansalliskirjaston sanoma- ja aikakauslehtikokoelman suomenkielinen osakorpus, Kielipankki-versio [tekstikorpus].* Kielipankki. Saatavilla http://urn.fi/urn:nbn:fi:lb-2016050302