



# Annif <3 Yle 2.0

*Annifin osittainen käyttöönotto artikkeleiden koneavusteisessa asiasanoituksessa*






Irene Nikkarinen  
Data Scientist, Sisältömetadatiimi








# Asiasanoitus Ylellä

Sisältöä asiasanoitettu jo jonkin aikaa koneavusteisesti käyttäen Leikiä

**Ulkomaiset sarjat** | Näytä kaikki

 <p>UUSI SARJA</p> <p>PHILHARMONIA</p> <p><b>Kapellimestari</b> Kapellimestari ristiriitojen keskellä</p>	 <p>KAUSI 3</p> <p>BABYLON BERLIN</p> <p><b>Babylon Berlin</b> Jännitystä suurkaupungin kuohuissa</p>	 <p>UUSI KAUSI</p> <p>BERLIN STATION</p> <p><b>Berlin Station</b> CIA-agentti saa soluttautumistehtävän</p>	 <p>SUOSIKKISARJA</p> <p>NORMAL PEOPLE</p> <p><b>Normaaleja ihmisiä</b> Hittikirjaan perustuva ihmishuhdraama</p>	 <p>UUSI SARJA</p> <p>FREUD</p> <p><b>Freud - tappajan mieli</b> Murhatutkintaa Freudin opeir</p>
--	--	--	--	---

**Kovat dokkarit** | Näytä kaikki

 <p>UUSI SARJA</p> <p>RANKKA VUOSI</p> <p><b>Rankka vuosi</b> Dokumenttisarja poikkeusajasta</p>	 <p>BLACK BOX SYRIA</p> <p><b>Syriän loputon sota</b> Miljoonien pako omasta maastaan, miksi?</p>	 <p>THE JUMP</p> <p><b>The Jump</b> Hyppy vapauteen</p>	 <p>OLEN HYVÄ ÄITI, KOSKA ANNOIN LAPSENI POIS</p> <p><b>Perjantai-dokkari</b> Olen hyvä äiti, koska annoin lapseni pois</p>	 <p>UUSI SARJA</p> <p>MIST SÄ TUUT?</p> <p><b>Mist sä tuut?</b> Sukellus suomirapin kaupunkelhin</p>
---	--	--	--	--

 <p>URHEILUTAPAHTUMAT</p> <p><b>Järjestäjä naukuttui apinoiksi:</b> Osa vanhemmista suuttui koronarajoituksista kilpa-aerobicin kisoissa, kun ei saanut lippua katsomoon</p> <p>30 kommenttia</p>	 <p>KORONAVIRUS</p> <p><b>Viedäänkö siv-tason huippujoukkueilta harjoittelupaikat alta?</b> Helsingin, Espoon ja Vantaan uusilla koronalinjauksilla olisi rajut vaikutukset urheiluun: "Lamaannuttaa toiminnan"</p>
 <p>MATKUSTUSRAJOITUKSET</p> <p><b>Islanti tarjoaa varakkaille etätyöviisumia ja Englannissa karanteeni lyhenee maksullisella koronatestillä – rajoitukset taipuvat paksulla lompakolla</b></p>	



# Mikä tekee Ylen käyttötapauksesta haastavan?

## *Asiasanaston laadussa on ongelmia*

- Paljon osittain tai täysin päällekkäisiä asiasanoja
  - Asiasanastossa syntyy luontaisesti hierarkia, jota emme tällä hetkellä mallinna
    - Asiasanat tuotu eri lähteistä ja ontologioista (Leiki, Wikidata, KOKO)
- Asiasanastoon lisätään uusia asiasanoja noin 200 per viikko
  - Duplikaattien siltaaminen käsityötä

['vesillä sattuneet onnettomuudet',  
'vesiliikenneonnettomuudet',  
'veneilyonnettomuudet']

['palo- ja pelastustoiminta', 'palo-  
ja pelastustyö', 'pelastuspalvelu',  
'pelastustoimi', 'pelastustoiminta',  
'palotoimi']

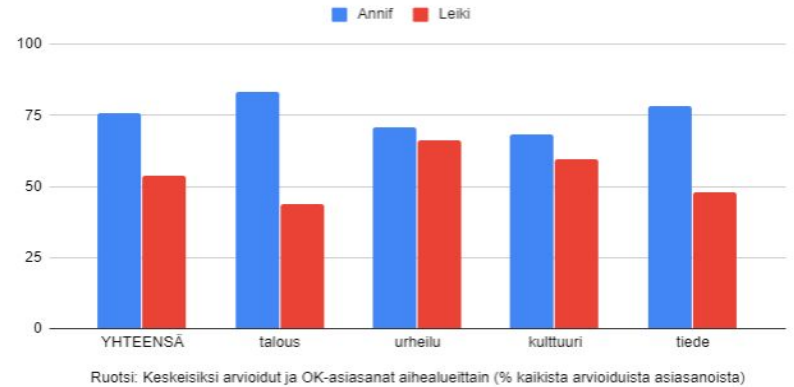
['pohjalaisia', 'pohjalainen',  
'pohjalaiset']

# Annifin ja Ylen historia

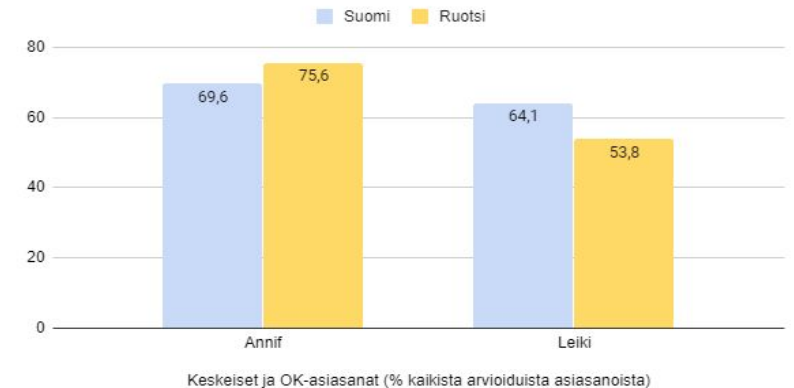
## Annifin ensitestit keväällä 2020

- Annifista Leikin korvaaja?
- Alkuperäisessä kokeilussa yhteensä noin 30 arvioijaa ja noin 100 artikkelia per kieli
  - Annif vaikutti toimivan vähintään yhtä hyvin kuin Leiki suurimmalla osalla aihealueista
- Edelleen käytössä silloin hyväksi havaitut backendit
  - Omikuji Attention
  - Maui
  - Yhdistetään nn ensemblella
    - Omikujin paino 2, Mauin 1

Keskeiset ja OK-asiasanat



Keskeiset ja OK-asiasanat



# Seuraavaksi laajempi testi

## Miten Annif toimii sisällöntuottajan arjessa?

- Otetaan Annif ensin käyttöön rajatussa joukossa julkaisujärjestelmiä
  - Mahdollisuus vertailla Leikiin
  - Konkreettisia lukuja joilla perustella Annifin käyttöönotto
- Ensimmäiseksi testijoukouksi valikoituivat artikkelijulkaisujärjestelmät SYND ja FYND
  - FYND tuottaa suomenkielistä aihesivustoa <https://yle.fi/aihe>
    - Aihesivujen tyyli ja aiheet eroavat jonkin verran uutis- ja ajankohtaistoimituksen sisällöstä
  - SYND tuottaa kaiken Ylen ruotsinkielisen sisällön: <https://svenska.yle.fi/>

na Urhellu Valikko

**Hoikka artisti, timmi artisti, laiha artisti – Nouseeko Suomen musiikkibisneksessä seinä vastaan, jos popmaailmaan yrittää lihavana laulajana?**  
12.03.2021 VLEX  
Poptähdet vaikuttavat olevan säännönmukaisesti hoikkia, timmejä tai laihoja. Miksi lihavana poptähtiä ei Suomessa juuri näy?

**Strömsö testaa tv-ohjelmien automaattista osiointia tekoälyn avulla**  
11.03.2021 STRÖMSÖ  
Osallistu tutkimushankkeeseen kokeilemalla uutta palvelua ja kertomalla meille kokemuksistasi.

Svenska.yle.fi

Coronavirus

Inrikes

Utrikes

Kultur

Sport

Huvudstadsregionen

Västnyland

Äboland

Österbotten

Östnyland

**NHL-kolumnen: Dags att sluta tvivla på vart Jesperi Kotkanemi är på väg som spelare - stortalangen kommer att bli en elitcenter i världens bästa liga**

Raseborgs stad förbereder sig för att ta över vid stängningshotade boende i Karis - oklart om värden kan fortsätta i samma lokaler

Boendet hotas stänga om Attendo inte anställer mer personal.

# Tutkimusasetelma

## *Miten saadaan parempaa tietoa Annifin toiminnasta?*

1. Testataan Annifia ensin SYND/FYNDissä
  2. Tallennetaan asiasanaehdotukset sekä Leikiltä että Annifilta
  3. Tarjoillaan toimittajalle Annifin asiasanat (n. 15 ehdotusta)
  4. Tarkistetaan kerran päivässä molempien mallien asiasanoille, mitkä tarjotuista asiasanoista on otettu julkaistuissa artikkelissa käyttöön
    - a. Haetaan tällöin ehdotukset molemmilta malleilta myös sellaisille artikkeleille joille ei haettu aikaisemmin
    - b. Tallennetaan analytiikka
- Asetelma on Annifille edullinen, sillä sisällöntuottaja näkee vain Annifin ehdotukset
  - Artikkelin ehdotusten ja lopullisten asiasanojen yhdistämistä ei voida tehdä täysin luotettavasti
    - Tehdään otsikon perusteella
  - Annif koulutetaan viikottain AWS:ssä käyttäen Ylen artikkelidataa

yle

Tuloksia





# Kerätty analytiikka

*Dataa Annifin käyttäytymisestä kerättiin noin kolme kuukautta*

	Ihminen <b>varmasti</b> nähnyt Annifin ehdotukset	Ihminen <b>mahdollisesti</b> nähnyt Annifin ehdotukset	Yhteensä
Suomi	114	337	451
Ruotsi	1173	2406	3579
Yhteensä	1287	2743	4030

# Kerätty analyysi

Dataa Annifin käyttäytyä

Näiden kahden sarakkeen väriero syntyy tilanteista, joissa artikkelin otsikkoa on muutettu asiasanaehdotusten hakemisen jälkeen

me kuukautta

	Ihminen <b>varmasti</b> nähnyt Annifin ehdotukset	Ihminen <b>mahdollisesti</b> nähnyt Annifin ehdotukset	Yhteensä
Suomi	114	337	451
Ruotsi	1173	2406	3579
Yhteensä	1287	2743	4030

# Paljonko Annifin ehdotuksia valitaan?

*Artikkeleille, joille ihminen on varmasti nähnyt Annifin ehdotukset*

malli	kieli	montako prosenttia asiansanoista mallin ehdotuksia keskiarvo	asiasanamäärä artikkelissa keskiarvo	valittujen ehdotusten keskiarvo	lisättyjen asiansanojen (ei ehdotettu mutta julkaistu) keskiarvo	artikkeleita
annif	sv	87%	10.13	8.70	1.33	1173
annif	fi	72%	10.46	7.27	3.66	114
leiki	sv	45%	10.13	4.46	5.53	1173
leiki	fi	39%	10.46	3.93	6.36	114

# Mille asiasanoille Annif toimii?

## *Annifin ja Leikin vertailua asiasanakohtaisesti*

- Asiasanoista ei juurikaan luotettavaa tietoa → analyysi jouduttiin tekemään pitkälti käsin
  - Mikäli asiasanoista tunnettaisiin hierarkia, voitaisiin tutkia aihekohtaisesti
- Laskettiin jokaiselle asiasanalle Leikin ja Annifin onnistumisprosentin erotus, ja tutkittiin asiasanoja jossa erot olivat suurimmat
- Mallit ehdottaneet kokeilussa vain n. 5% kaikista asiasanoista
  - Ei voida tehdä lopullisia päätelmiä

- Leiki saattaa toimia paremmin erisnimissä
  - Urheilu ja urheilijat
- Asiasanat joissa Annif on toiminut Leikiä paremmin ovat pitkälti yleiskäsitteitä

Pohjois-Lapin seutukunta
Kysyntä ja tarjonta
purjehdus
virkamiehet
lahjonta
nopeusrajoitukset
kevyt liikenne
rikosilmoitukset
ammattilliset oppilaitokset
lahjoitukset
teleoperaattorit
julkisuuden henkilöt
asuntopalot
ekoenergia
siivous
mellakat
taloudelliset ennusteet

Asiasanat, joille **Annifin** ennustukset ovat valittu useimmin verrattuna Leikiin (asiasanoista, joita molemmat mallit ovat ennustaneet)

Anaheim Ducks
Nigeria
markkinat
viinit
Saaristomeri
Kilpailu- ja kuluttajavirasto
Nepal
syntymäpäivät
Mika Lintilä
Kiekko-Vantaa
Linda Sällström
pseudotiede
apartheid
Martin Luther King, Jr.
Sex Pistols
Nikolaj Ehlers
Alexandria Ocasio-Cortez

Asiasanat, joille **Leikin** ennustukset ovat valittu useimmin verrattuna Leikiin (asiasanoista, joita molemmat mallit ovat ennustaneet)



# Mitä asiasanoja lisätään eniten?

Suosittuja asiasanoja, koska luku on absoluuttinen

Lisäksi paljon ajankohtaisiasiasanoja

- Virustaudit
- Marinin hallitus
- Capitolin valtaus

COVID-19
koronavirus
Kotimaan uutiset
Pohjanmaa
Koronarokote
uutiset
Ulkomaat
kulttuuri
sää
sääennusteet
Euroopan unioni
Niko Kytösaho
tietokilpailut
Keski-Pohjanmaa

Annif

Kotimaan uutiset
Ulkomaat
urheilu
koronavirus
COVID-19
tartuntataudit
kulttuuri
talous
rokotus
Yhdysvallat
virustaudit
Koronarokote
Vaasa
jääkiekkoilijat

Leiki

# Mitä heikkouksia Annifilla on?

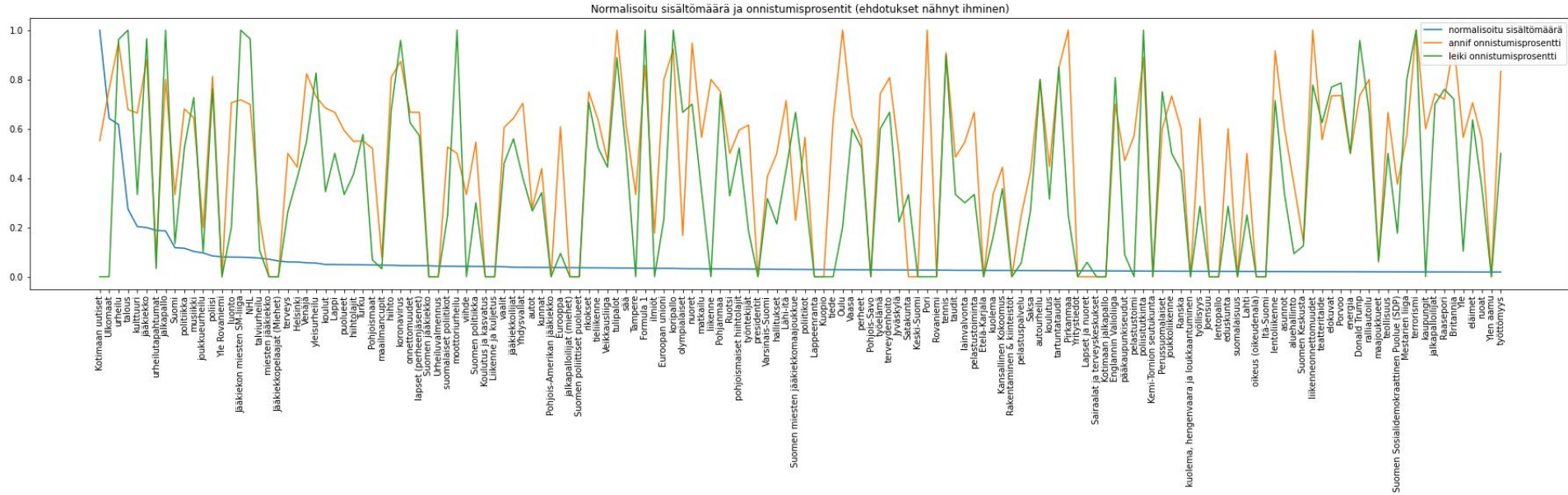
- Annif assosioi asiasanoja, ja saattaa ehdottaa väärää, mutta saman aihealueen asiasanaa
  - *Kerttu Niskanen* tarjotaan asiasanaksi artikkeliin Riitta-Liisa Roposesta
  - *Joulu* ehdotetaan asiasanaksi jalkapalloa käsitteleviin artikkeleihin
    - Uudelleenkoulutus niin, että kannustetaan ennustamaan asiasanoja, jotka esiintyvät yhdessä?
- Oleellisia asiasanoja ehdotetaan matalalla relevanssilla





# Annifin ja Leikin onnistumisprosentit

...yhdessä asiasanan normalisoidun sisältömäärän kanssa



# Ennustetaanko yleisiä asiasanoja paremmin?

*Korrelaatioita sisältömäärän kanssa*

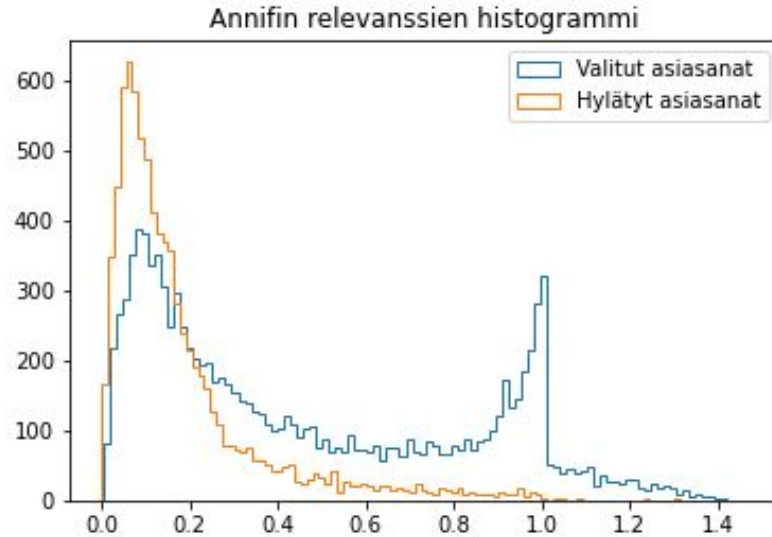
Muuttujat	Malli	Spearman-korrelaatio	p-arvo
Onnistumisprosentti ja sisältömäärä	Annif	0.34	< 0.01
Onnistumisprosentti ja sisältömäärä	Leiki	0.36	< 0.01
Ehdotuskerrat ja sisältömäärä	Annif	0.46	< 0.01
Ehdotuskerrat ja sisältömäärä	Leiki	0.57	< 0.01

Käyttäen artikkeleita, jotka on varmasti nähnyt ihminen

Mallin korrelaatiota laskettaessa käytetty vain sellaisia asiasanoja, joita malli on ennustanut kokeilun aikana

# Annifin relevanssit

*Ovatko valittujen asiasanojen relevanssit suuremmat kuin hylättyjen?*

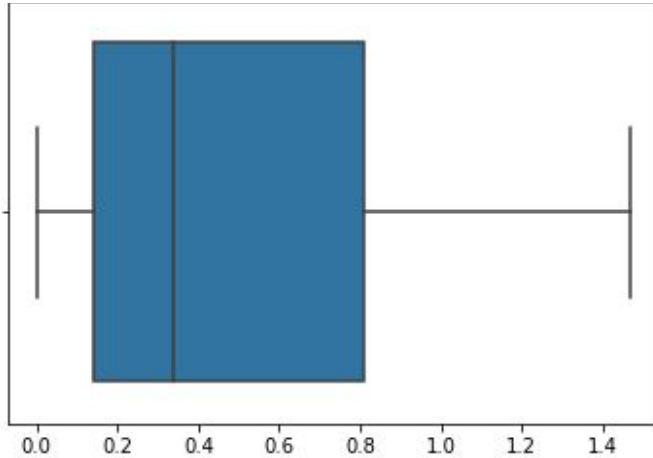


Käytetty artikkelia, jotka varmasti nähnyt ihminen

# Annifin relevanssit

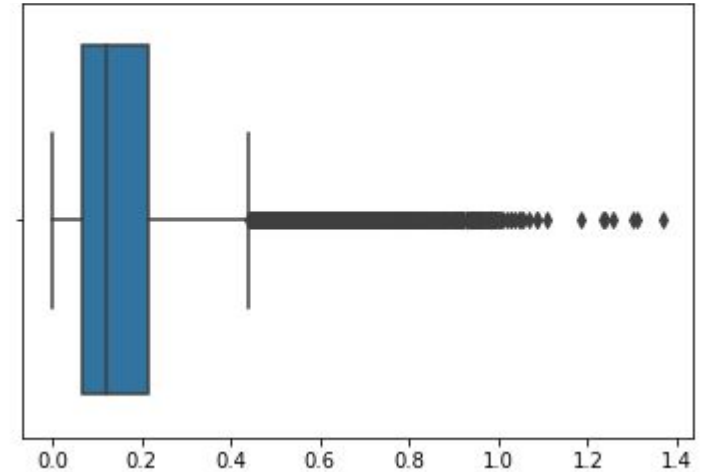
*Ovatko valittujen asiasanojen relevanssit suuremmat kuin hylättyjen?*

Boxplot **valittujen** asiasanojen relevansseista molemmille kielille



Käytetty artikkelieita, jotka varmasti nähnyt ihminen

Boxplot **hylättyjen** asiasanojen relevansseista molemmille kielille





# Lopputulemat

*Mikä on Annifin tulevaisuus Ylellä?*



# Annifin tulokset ovat olleet lupaavia

## *Annif tarjoaa Ylelle paljon mahdollisuuksia*

- SYND:in ja FYND:in sisällöntuottajilta ei ole tullut valituksia Annifin ehdotuksista
  - No news is good news
- Annifia ollaan myös testattu asiasanoittamaan ohjelmia niiden litteraatioiden perusteella
  - Tulokset positiivisia
  - Tulokset saattavat parantua kouluttamalla litteraatioilla
- Mahdollisuus kouluttaa Annif erikoistumaan erilaisiin sisältöihin
- Tarjottujen ehdotusten rajaaminen
  - Onko tärkeämpää, ettei hyviä asiasanoja leikkaudu pois, vai ettei niitä ole ylimääräisiä?
- Annifin ennustusten laatu sidoksissa Ylen asianaston laatuun
  - Annifille tarjottuja luokkia rajoitettava
- Leikissä mahdollisuus painottaa otsikkoa -- sama toiminnallisuus Annifin?

# Kiitos! Tack!

[irene.nikkarinen@yle.fi](mailto:irene.nikkarinen@yle.fi)