JYU REPORTS 15

Markus Mykkänen & Aleksi Koski

JOURNALISTISEN TIEDONHANKINNAN TEHOSTAMINEN TEKSTIANALYYSITYÖKALUILLA

Tutkimushankkeen loppuraportti ja ohjelmistoarvioinnit toimituskäyttöön

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYU REPORTS 15

Markus Mykkänen & Aleksi Koski

JOURNALISTISEN TIEDONHANKINNAN TEHOSTAMINEN TEKSTIANALYYSITYÖKALUILLA

Tutkimushankkeen loppuraportti ja ohjelmistoarvioinnit toimituskäyttöön

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2022

Permanent link to this publication: http://urn.fi/URN:ISBN: 978-951-39-9432-7

ISBN (PDF) 978-951-39-9432-7
URN:ISBN:978-951-39-9432-7
ISSN 2737-0046
DOI: 10.17011/jyureports/2022/15

Sisällys

ESIPUHE

Journalisteilla voi uutistyön taustalla olla toisinaan suuri aineisto. Taustamateriaalit voivat käsittää jopa tuhansia sivuja kymmenistä eri lähteistä. Toimittajille voi olla haastavaa tai ylivoimaista nähdä aineistosta, mikä olisi mielenkiintoisinta tai uutisen arvoista. Eri alojen tutkijat ovat luoneet monenlaisia tekstin louhinta-, laskenta- ja analyysiohjelmia tutkimustyöhön isojen asiakirja-aineistojen läpikäymiseen. Heillä on näihin perinpohjaisia lähestymistapoja, joilla saa aikaiseksi vankasti perusteltuja tutkimustuloksia. Toimittajien tiedontarpeet ovat kuitenkin erilaisia. Yksittäisistä uutiskriteerit täyttävistä löydöistä voi uutisoida välittömästi löydön jälkeen. Aineistoon voi palata uudestaan myöhemmin kattavampaa tarkastelua varten. Toimittaja voi käyttää omiin aineistoihinsa tutkijoiden ohjelmia ja löytää helmiä muutamalla klikkauksella. Tässä hankkeessa etsittiin ratkaisuja tähän ongelmaan eli kuinka nopeasti selata läpi aineisto ja löytää uutisarvoinen tieto? Mikä on paras työväline tähän?

Markus Mykkänen ja Aleksi Koski

1 TAUSTA

Toimitukset laativat uutiset oman tiedonhankinnan pohjalta. Tavallisesti journalistin tärkein resurssi on aika. Useimmilla journalisteilla ei ole erityiskoulutusta tiedonhankintaan tai informaatioteknologiaan, eikä heillä ole rahaa tai aikaa paneutua asiaan. Tietoa halutaan mahdollisimman runsaasti ja hyödynnettäväksi valitaan usein ensimmäiseksi kohdalle osuva, riittävän uutisarvoiseksi koettu tieto. Syynä tähän ovat erityisesti journalististen medioiden kustannuspaineet. (Attfield & Dowell, 2003; Cambell 1997; Hopeakunnas 2015)

Journalistisia tiedonhankintavälineitä ovat esimerkiksi aiemmin julkaistut tiedot (uutiset ja tiedotteet), julkaisemattomat viranomaisasiakirjat sekä suorat henkilölähteet eli haastattelut. Erityisesti tutkiva journalismi edellyttää perinpohjaista perehtymistä asiakirjalähteisiin. Niistä voi löytyä tietoja, joita viranomainen pyrkii piilottamaan tai joilla voi esimerkiksi tarkastaa ja kumota virkamiesten tai poliitikkojen haastatteluissa esittämät väitteet. Viranomaisten asiakirjatiedot paljastavat sitä, mitä viranomaiset tekevät, sen sijaan mitä he haluaisivat kertoa tekevänsä. (Houston & IRE, 2008).

Toimittajien syyt olla käyttämättä asiakirjalähteitä liittyvät tavallisesti aikaan: asiakirjojen saaminen vie aikaa, niiden läpikäyminen vie aikaa, eikä ole takeita, että niistä löytyy uutista. Moni toimittaja käyttää vain tiedoteaineistoa ja haastatteluja, koska ne riittävät uutisia varten. Henkilölähteet voivat kuitenkin antaa haastatteluja myös muille medioille ja myös tiedotteet lähetetään kaikille medioille. Useat viime aikojen valtakunnallisista skuupeista pohjaavatkin asiakirjatietoihin, joiden kaivelemiseen muut mediat eivät ole käyttäneet aikaa (esimerkiksi Harjumaa & Jansson, 2020; Paananen & Liski, 2020; Rämö & Liski 2018).

Tiedonhankinta asiakirjoista käsittää useita huomioitavia seikkoja: lähteiden etsimisen, tiedon hankinnan tietokannoista tai tietopyynnöillä viranomaislähteistä, sekä käsittelyn, lukemisen ja läpikäymisen. Viranomaisten asiakirjatietojen tiedonhankinnasta on ollut Jyväskylän yliopistossa erilaisia tutkimushankkeita aiempina vuosina. (julkisuuslaki.fi)

Asiakirjojen käsittelyyn on olemassa koneavusteisia lähestymistapoja, kuten erilaiset algoritmit, koneoppiminen ja scraping-tekniikat. Näistä useimmat edellyttävät ohjelmointiosaamista, jota lähtökohtaisesti useimmilla toimittajilla ei ole. Moni menetelmä edellyttää joko lingvistiikan tai tietotekniikan syventävää opiskelua. Ei voida olettaa kiireisen journalistin ryhtyvän vuosien opiskelu-urakkaan käyttääkseen uutta tiedonhankintametodia (Bednarek & Carr, 2020; De Grove, Boghe & De Marez, 2020; Hase, Engelke & Kieslich, 2020; Su, Hu & Lee, 2020; Heyl, Joubert & Guenther, 2020).

Toimittajien hyödynnettävissä on myös korpuslingvistiikan kentältä peräisin olevia tekstianalyysityökaluja, joiden käyttö on nopeaa, helppoa ja intuitiivista. Toimitusten ja toimittajien tekstianalyysityökalujen käyttöönottoa voidaan edistää lyhyillä, muutaman minuutin screencast-opasvideoilla, joilla opastetaan ohjelmien ja niiden ominaisuuksien käyttöä. Koneavusteisen tekstianalyysin tutkiminen osana journalistista työprosessia on vielä uutta. Eri tutkijat ovat käyttäneet hyvin erilaisia välineitä ja menetelmiä, eikä ole täysin ilmeistä, mitä kaikkea näillä uusilla välineillä voisi tehdä. Toimituksiin tarvitaan menetelmien yleistajuista tietoa ja osaamista. (Bednarek & Carr, 2020; De Grove, Boghe & De Marez, 2020).

Tutkimuksen menetelmät

Tämä tutkimus toteutettiin 1.9.2021–31.8.2022 osapäiväisenä tutkimus- ja kehitystyönä. Hankkeen alussa kirjallisuuskatsauksella selvitettiin journalistisessa tutkimuksessa hyödynnetyt tekstianalyysiohjelmat ja -menetelmät. Näistä valittiin vertailuun sellaiset, jotka vastasivat journalistisen työn ominaisuuksia, tavoitteita ja reunaehtoja. Kirjallisuuskatsauksen myötä tarkentuivat myös ohjelmistojen arvioinnin kriteeristöt.

Journalismin ideologisista kriteereistä huomioitaviksi tulivat riippumattomuuden, objektiivisuuteen pyrkimisen, lähdesuojan sekä tietosuojan kysymykset (Deuze, 2005; Pöyhtäri, Väliverronen & Ahva 2016):

- Ohjelman on oltava mielellään maksuton tai hyvin edullinen. Tämä vaatimus karsi eniten erilaisia ohjelmia.
- Toimittajien on aina kyettävä pysymään perillä siitä, mitä he ovat tehneet ja miksi. Mahdollisimman vähän tapahtuu piilossa ohjelman sisällä. Kirjallisuuskatsauksessa erottui erilaiset algoritmeihin ja esimerkiksi koneoppimiseen painottuvat lähestymistavat, jotka karsittiin tästä tutkimuksesta pois: Tällaisten ohjelmien käyttö ilman käsitystä siitä miten ohjelmat tuloksensa saavat aikaiseksi voisi olla epäeettinen lähestymistapa.
- Ohjelman on oltava täysin toimittajan hallinnassa tämän omilla laitteilla, eikä yhteydessä pilvipalveluun jonne tiedot ovat vaarassa latautua. Erityisesti moni uudempi niin sanottua ”big datan” tutkimusta mahdollistava lähestymistapa karsiutui tätä myötä pois. Toisaalta tällaiset vaikuttivat edellyttävän syötetyltä datalta puhtautta ja yhdenmukaisuutta rakenteen kannalta, joka ei jokapäiväisessä uutistyössä ole mahdollista.

Journalismin ideologinen vaatimus on myös ajankohtaisuus eli uutisia on kyettävä tuottamaan päivänpolttavista asioista nopeasti. Tästä tavoitteesta syntyy keskeisiä käytännöllisiä reunaehtoja ohjelmien käytölle, koska keskeisin resurssi on aika. Toimittajien aiemman koulutuksen ja osaamisen puute karsi tehokkaasti esimerkiksi koodausta tai erityiskoulutusta edellyttäviä ohjelmia. Käytännössä jäljelle jää vahvasti visuaaliset käyttöliittymät, joissa on varsin yksinkertaiset toiminnot. Käytännön journalistisen työn vaatimuksia olivat siis (muun muassa Brehmer ym., 2014 ja Liu ym., 2013 pohjalta):

- Ohjelmien käynnistyksen ja käytön on oltava nopeaa ja sujuvaa,
- Ohjelman on oltava visuaalinen ja helppokäyttöinen, eikä vaadi merkittävää ennakko-opettelua,
- Ohjelman on tuettava mahdollisimman laajaa skaalaa eri tiedostomuotoja, joita on voitava ladata ohjelmaan kerralla massoittain (robust import).
- Ohjelman on voitava käsitellä hyvinkin “likaista dataa”, eli sotkuista ja rakentumatonta tietoa (robust analysis).
- Luvut ja lukumäärät ovat raportoinnissa tärkeitä, näiden tuottaminen eri tavoin ja vaivattomasti on keskeistä.
- On tärkeää, että aineistoon voi vaivattomasti käydä käsiksi ja esimerkiksi tunnistaa, poimia ja lukea yksittäisiä asiakirjoja sekä tehdä niihin

tunnistamista helpottavia muistiinpanoja. Siirtymisen eri näkymien osalta täytyy olla siis selkeää ja nopeaa.

Valitun näkökulman oli tarkoitus vastata mahdollisten rivitoimittajien osaamistasoa. Vertailusta luodusta raportista sekä ohjelmien arvosteluista journalistit voivat arvioida nykyisten tekstianalyysiohjelmien käyttökelpoisuutta työssään. Tämän lisäksi toimittajille laadittiin opas ja screencast-videoita rivitoimittajan tarpeille parhaaksi valikoidusta ohjelmasta.

Tutkimuksen kulku

Hanke lähti liikkeelle vertaisarvioidusta tutkimusartikkelista (Bednarek & Carr 2020), jossa esiteltiin tekstianalyysiohjelmia tiedonhankintamenetelmänä. Hankkeen aluksi toteutettiin kirjallisuuskatsaus, jota ohjasivat seuraavat tutkimuskysymykset:

- Millaisia tekstianalyysityökaluja journalismin tutkimuksessa on käytetty?
- Millaisin menetelmin näitä työkaluja on käytetty?

Kirjallisuuskatsaus tuotti hakukriteerien pohjalta joko hyvin suppean tai liian laajan tuloksen: lähestymistapa on journalistiikan tutkimuksessa hyvin uusi, ja useiden tieteenalojen alalta ohjelmien käyttö on kirjavaa. Ohjelmien käyttö tutkimuksessa ja erityisesti näiden raportointi on hyvin vaihtelevaa. Kirjallisuuskatsausta laajennettiin ns. snowballing-menetelmällä seuraten katsauksessa löydetyn kirjallisuuden viittauksia sekä kirjallisuuteen viitannutta uutta kirjallisuutta. Kirjallisuutta, menetelmiä ja työkaluja kerättiin taulukkoon.

Tutkimushankkeen näkökulmaa rajasi edeltävässä luvussa esitelty journalistisen työn näkökulma sekä alkuperäisen artikkelin laskennallinen ja nopeaan hakuun ja tekstimassojen laskennalliseen analyysiin perustunut lähestymistapa. Tämä rajasi tutkijoiden hakemien ohjelmien vaatimuksia ja ominaisuuksia. Valikoituneissa tutkimusartikkeleissa mainitut ohjelmat ovat pääsääntöisesti korpuslingvistiikkaa varten alun perin kehitettyjä. Valitsemamme ohjelmat ovat myös lähteiden valossa yleisimmin maailmalla tarkoitukseensa käytettyjä. Vähemmän käytetyistä ohjelmista ei löytynyt kilpaili-

joita valituille ohjelmille: kaikkiaan erilaisia ohjelmia ja oikeastaan usein lähinnä apuohjelmiksi laskettavia toiminnaltaan rajallisia ohjelmia kokeiltiin kymmenittäin. Valtaosa kuitenkin karsittiin nopeasti lähinnä korkean hinnan, liian suuren oppimiskynnyksen tai erittäin kevyen toiminnallisuuden vuoksi.

Kirjallisuuskatsauksen sekä ohjelmien kokeilun jälkeen tehtiin laajemmat käyttökokeilut neljällä valitulla ohjelmalla. Ohjelmat asennettiin ja niiden käyttöä opeteltiin riittävästi, että niillä kyettiin tekemään testihaut kirjallisuuden pohjalta valituilla kolmella eri lähestymistavalla. Lähestymistavoiksi valittiin kirjallisuudessa mainitut perustoiminnallisuudet, joiden käyttö ei edellytä erityistä ennakko-osaamista, mutta mikä antaa merkittävää etua tiedonhankinnassa. Näitä ovat: 1. sanalistojen luonti ja sanojen laskeminen, 2. konkordanssihaku, 3. monipuoliset ja sujuvat hakumahdollisuudet halutuille avainsanoille. Tässä tutkimushankkeen ensimmäisessä vaiheessa päädyttiin keskittymään yksinkertaisiin, itsestään selviin ja toimittajien muutoinkin tiedonhankinnassa käyttämiin menetelmiin. Hankkeen toisessa, myöhemmin toteutettavassa vaiheessa katsotaan toimittajien kanssa, miten paljon erilaisia menetelmiä voidaan laajentaa ja tuoko se lisäarvoa vai kannattaako pysyttäytyä yksinkertaisissa perustoiminnallisuuksissa.

Aineistohaut tehtiin englanninkielisellä sekä suomenkielisellä aineistolla. Englanninkielinen aineisto käsitti kirjallisuuskatsauksen tuloksen, suomenkielinen aineisto hallinto-oikeuksien ratkaisuja .pdf-tiedostoina. Molemmissa aineistoissa etuna oli se, että .pdf-tiedostot oli toteutettu osin suorina skannauksina kuviksi eikä tekstiksi, ja erityisesti englanninkielisessä aineistossa asiakirjojen rakenne oli hyvin vaihtelevaa johtuen lukuisista eri julkaisijoista.

Arviointi itsessään oli laadullista tulkintaa, jossa raportoitiin havainnot ohjelmien käytöstä kuhunkin tehtävään huomioiden samalla alussa mainitut journalistisen työn vaatimukset. Arvostelut ovat luettavissa seuraavissa luvuissa. Niissä pääpaino on kunkin ohjelman erityislaadun sekä erilaisten käyttötarkoitusten esittelemisessä, erilaisissa tilanteissa eri ohjelma voikin olla parempi.

2 TULOKSET

Hankkeessa tarkemmin arvioitujen neljän lupaavimman ohjelman vertailun yhteenveto tässä:

Ominaisuus	**1. sija**	**2. sija**	**3. sija**	**4. sija**
Asennuksen helppous	Voyant Tools/ AntConc	WordSmith		Overview
Käynnistys-nopeus (aineiston-lataus)	Voyant Tools	WordSmith	AntConc	Overview
Monipuolinen aineistonlataus	Over-view	Voyant Tools	AntConc	WordSmith
Monipuolinen analyysi / Toi-minnallisuuksien monipuolisuus	Voyant Tools	AntConc	Overview	WordSmith
Monipuolinen vienti	AntConc	Wordsmith	Voyant Tools	Overview
Hakujen toimi-vuus/käyttö	Voyant Tools	AntConc	WordSmith	Overview
Asetusten tallentaminen/ ohjelman räätälöinti	Ant-Conc/ Over-view	WordSmith		Voyant Tools

Taulukko 1. Yhteenveto neljästä arvioidusta tekstianalyysityökalusta.

Asennuksen helppous

Voyant Tools toimii heti selaimessa. Mutta jos haluaa kaiken aineiston pysyvän omalla koneella, on käynnistettävä Java-pohjainen palvelin. Tätä helpompaa on toki asentaa AntConc ja käynnistää se normaalisti. Wordsmith on muuten yhtä helppo, mutta oikean version löytäminen ja rekisteröiminen ilmaiskoodilla vaatii tietoa. Overview taas ei käynnisty ilman erillistä virtuaaliympäristöä Windowsissa.

Käynnistysnopeus

Voyant Tools kykenee lataamaan suurenkin aineiston nopeimmin. Wordsmith on myös nopea, mutta edellyttää aineiston muuntamista tekstitiedostoiksi. AntConc ei edellytä aineiston muuntamista, mutta suorittaa kauan suuremman aineiston latausta: noin kahden gigatavun aineistoa Voyant Tools mietti testikoneella 1,5 minuuttia, AntConcilla tunnin. Overview tekee aineiston lataamisen nopeammin, mutta itse ohjelman käynnistys on mutkikasta virtuaaliympäristön ja virheilmoitusten vuoksi.

Monipuolinen aineistolataus:

Overview tukee laajinta skaalaa erilaisia tiedostomuotoja, jopa Powerpoint-tiedostoja. Voyant Tools tulee toisena ja myös AntConcin tiedostotuki on laaja, mutta se ei tue esimerkiksi MS Word -tiedostojen suoraa aineistolatausta. Wordsmith edellyttää kaiken aineiston muuntamista tekstitiedostoiksi (.txt).

Monipuolinen analyysi / Toiminnallisuuksien monipuolisuus:

Voyant Toolsissa on eniten erilaisia työkaluja, eli erilaisia tapoja analysoida tietoa. AntConc tulee hyvänä kakkosena ja on ehkä tuloksissaan kaikkein kattavin, eli sen kaikki työkalut ovat niin sanotusti kovia työkaluja, eivät visuaalisia orientaatioapuja. Overviewn valikoima on monipuolinen mutta melko suppea.Wordsmithissä vaihtoehdot ovat kaikista niukimmat ja visuaaliset avut puuttuvat tyystin.

Monipuolinen vienti

Kaikkien analysoitujen ohjelmien vienti eli tulosten tallentaminen muiden ohjelmien käyttämiin tiedostomuotoihin on melko vaatimaton. AntConcista nämä ominaisuudet puuttuivat kauan ja on aika vastikään vasta lisätty, kattaen esimerkiksi teksti-, html- ja Microsoft Excel-tiedostot. Wordsmith 5.0

tarjoaa teksti-, XML- ja Excel-tiedostoja. Voyant Tools tarjoaa niukempaa keinovalikoimaa ja selkeimmät tavat ovat .csv-tiedosto (tekstitiedosto, jota esimerkiksi Excel osaa lukea) sekä suora copy/paste. Overviewssa on vain Excel-toiminto.

Hakujen toimivuus/käyttö

Kaikissa työkaluissa aineistosta hakeminen avainsanoilla on nopeaa ja sujuvaa. Sanalistat löytyvät, konkordanssihaku toimii ja asiasanalla etsiminen onnistuu. Voyant Tools vetää pisimmän korren käytön nopeudessa ja vaivattomuudessa, kun useat näkymät ovat auki yhtä aikaa. AntConc on kulmikkaampi, mutta tuottaa kattavia tuloksia, Wordsmith samoin mutta suppeammalla keinovalikoimalla. Overviewn tulokset olivat heikoimmat: datan likaisuus eli tiedon rakenteen epäselvyydet tai virheet sotkevat sen toimintaa selvästi eniten ja haut painottavat yleisimpiä tuloksia, jolloin on vaikea löytää harvempia osumia saaneita sanoja.

Asetusten tallentaminen/ohjelman räätälöinti

Voyant Toolsin asetuksia ei voi tallentaa. Tällöin räätälöinti jää erillisten asetusten tallentamiseen tekstitiedostoille. Muilla ohjelmilla taas tallennetut asetukset säilyvät seuraavaa käynnistyskertaa varten ja ne voi myös tallentaa ja jakaa yheistyökumppaneille tai varmuuskopiona. AntConc, Overview ja Voyant Tools mahdollistavat myös avoimen lähdekoodin projekteina koodaritaitoiselle mahdollisuuksia säätää ohjelmaa omilla tavoilla.

Yhteenveto:

Tutkimushankkeen tavoitteena oli selvittää yksi ja paras tekstianalyysityökalu toimittajien työn tueksi, sekä laatia sen käytöstä mahdollisimman yksinkertainen opas ja ohjevideot. Täksi ohjelmaksi valikoitui Voyant Tools, koska se parhaiten täytti yleisjournalismin vaatimukset: se on hyvin intuitiivinen ja helppokäyttöinen ja sen kokeileminen ei vaadi minkäänlaista pohjaosaamista tai perehtymistä esimerkiksi korpuslingvistiikkaan. Ohjelma on silti hyvin monipuolinen ja kattava sekä ennen kaikkea hyvin nopea aineistonanalyysikeino. Se soveltuu hyvin vaikkapa jokapäiväiseen journalistiseen työkäyttöön aineiston analysoinnissa ja tutkimisessa.

Voyant Tools

Voyant Tools on lähtökohtaisesti web-pohjainen tekstin analyysiohjelma. Sen Java-pohjaisen serverin voi myös ladata omalle koneelle, jolloin ohjelma aukeaa edelleen oman selaimen kautta, mutta ei toimita siihen syötettyä aineistoa pilvipalveluun Yhdysvaltoihin.

Voyant Tools on helppo ja vaivaton käyttää: siihen voi ladata muun muassa teksti-, HTML-, XML-, PDF-, RTF- ja Microsoft Word -tiedostoja. Latauksen jälkeen esiin nousee useita erilaisia analyysi-ikkunoita, jotka ovat auki vierekkäin. Ikkunoiden kokoa, järjestystä ja sisältöä voi muuttaa, kaikki vaihtoehtoiset työkalut eivät ole heti nähtävillä.

Helppoa kuin heinänteko

Voyant Toolsin käyttö verkossa edellyttää vain vierailua verkkosivulla https://voyant-tools.org/. Esimerkiksi täysin julkisen aineiston kuten verkkosivuilta ladatun materiaalin voi huoletta ladata ohjelmaan ja katsoa, mitä kaikkea aineistosta paljastuu.

Mikäli ei halua aineistonsa päätyvän ulkomaalaiselle pilvipalvelimelle, täytyy luoda paikallinen serveri. Ohjeet ja linkit löytyvät osoitteesta https://voyant-tools.org/docs/#!/guide/server. Lisäksi paikallista serveriä varten pitää ladata Java-ohjelmistoympäristö. Javan saa ladattua osoitteesta https://www.java.com/en/download/

Voyant Toolsin suurin miinus on tämä paikallisen serverin Java-pohjaisuus, koska serverin käynnistäminen vie aina aikansa ja Java ohjelmistona täytyy päivittää melko usein. Ohjelman tekniset laitteistovaatimuksetkin ovat oikeastaan Javan käyttövaatimukset. Lisäksi Voyant Toolsissa ei ole mitään automaattista päivitystä, vaan uusi kehitetty versio on osattava hakea itse verkosta.

Kun aineisto on ladattu Voyant Toolsiin ja analyysinäkymä aukeaa, voi se ensin vaikuttaa sekavalta, koska erilaisia analyysinäkymiä on auki kerralla viisi kappaletta. Tämä on kuitenkin työkalun suurin etu: erilaisia näkymiä voi käsitellä yhtaikaa ja esimerkiksi yksittäisen asiakirjan tarkastelu onnistuu nopeasti työkalun avulla samanaikaisesti, kun muut ikkunat ovat auki. Lisäksi tehtävästä riippuen useimpia ikkunoita ei yleensä tarvita. Joskin erilaisia työkaluja on käytettävissä enemmän kuin mitä näkyvillä on näkymiä. Näkymät voi vaihtaa mieleisekseen.

Erilaisia työkaluja on 24 erilaista, joista tosin suurin osa on erilaisia visuaalisia apuvälineitä, joiden suora hyöty on vaatimaton. Erilaiset tiedonhankintatutkimukset kuitenkin korostavat, että varsinkin aineiston ensisilmäyksen tukena visuaalisilla työkaluilla voi olla suuri orientoiva merkitys.

Perustyön perustyökalu

Toinen miinus ohjelmassa on, että siinä ei ole omaa tallennusmahdollisuutta. Joka käynnistyskerralla se alkaa ikään kuin nollatilanteesta eli siihen ei voi tallentaa omia asetuksia. Mutta toisaalta, eipä näitä tarvitsekaan, sillä tulokset voi viedä esimerkiksi .csv-tiedostoiksi avattavaksi vaikkapa Microsoft Excelissä ja tarvittavia asetuksia voi pitää omalla tekstitiedostolla (.txt).

Voyant Tools on kaikista hankkeessa kokeilemistamme työkaluista selvästi helppotajuisin ja nopein käyttää menestyksekkäästi. Se sopii erinomaisesti suurten aineistojen selailuun ja etsimiseen toimittajan perustyön ohessa. Siksi se on valittu tässä hankkeessa opasvideoiden kohteeksi ja sen käytöstä on tarkempi opas. Isompiin ja pitempiin tutkiviin projekteihin voi sitten käyttää esimerkiksi AntConcia.

PERUSINFO:
Kehittäjä: Stéfan Sinclair & Geoffrey Rockwell
Julkaisuvuosi: 2003
Kehitysvaihe: Päivitetään edelleen, oheen luodaan uusia ohjelmia kuten yhteistyöohjelma Spyral
Hinta: Ilmainen
Laitevaatimukset: Java 8 tai uudempi, kovalevytilaa 5 Gt + 126 Mb (Java), keskusmuistia 128 Mb (Java)
Käyttöjärjestelmätuki: selainpohjainen - Windows, Linux ja Mac
Mobiilituki: toteutettu HTML5-pohjaisena, toimii mobiililaitteilla MUTTA paikallisen serverin pyörittäminen edellyttää Javaa.

Wordsmith

Wordsmith on vanha ja edelleen kehitettävä kielitieteilijöiden apuväline sanojen laskemiseen ja kielen rakenteiden tulkitsemiseen. Wordsmithin vahvuuksia ovat yksinkertaisuus ja selkeys, huonoja puolia ominaisuuksien puuttuminen ja toiminnan rajallisuus muihin vertailtuihin ohjelmiin verrattuna.

Ohjelmasta on vuosien mittaan luotu uusia versioita. Versiot 5.0 ja 4.0 ovat ladattavissa ilmaiseksi. Uusimmat versiot maksavat 50 puntaa plus arvonlisävero. Tämä arvostelu koskee versiota 5.0.

Yksinkertaista laskentaa

WordSmithin voi ladata osoitteesta https://www.lexically.net/wordsmith/downloads/. WordSmith 5.0 versioon on asennuslinkki sekä ilmaisen rekisteröimisavaimen tiedot osoitteessa https://lexically.net/wordsmith/version5/.

Lataus sekä asennus on helppoa ja nopeaa, ohjelma toimii itsenäisesti kotikoneella eikä ole yhteydessä verkkoon. Wordsmithin käyttö on selkeää ja perusnäkymä on yksinkertainen. Wordsmithin ytimessä on kolme ohjelmaa: WordList, joka laskee sanat aineistosta, KeyWords, joka vertaa aineistoa suurempaan vastaavaa kieltä olevaan tekstiaineistoon, sekä Concord, joka tekee valituista sanoista konkordanssihaun.

Ohjelman suurimpia huonoja puolia on eri tiedostomuotojen tuen puute. Ohjelma edellyttää aineiston muutamista tekstitiedostoiksi (.txt). Se onnistuu helpoiten esimerkiksi massakonversio-ohjelmalla, kuten Xpdfreader tools, joka osaa muuttaa myös kuvia tekstiksi (OCR). Tehdasasetukset ovat hyvät sanalistojen tekemiseen ja konkordanssiin. Asetukset ovat englanninkieliset, mutta sieltä kannattaa vaihtaa päälle suomen kielen tuki.

Avainsanojen etsiminen edellyttää verrokiksi suurempaa tekstiaineistoa, korpusta. Tämän tekstiaineiston pitäisi olla samantyylistä tekstiä kuin tutkittava aineisto. Wordsmithin valmistanut yritys Lexically viittaa itse sivuillaan tutkimukseen, jonka mukaan viisi kertaa suurempi aineisto on riittävä verrokkiaineisto avainsanojen etsimistä varten.

Konkordanssihaku toimii nopeasti ja sujuvasti. Mutta haun huono puoli on monipuolisuuden puute: erilaisia näkymiä ja varsinkin visuaalisia työkaluja puuttuu.

Kun tarkoitusperät on selkeät ja suppeat

Tutkijoiden kannalta Wordsmithin etuna on se, että sitä on käytetty paljon erilaisessa tutkimuksessa, joten siitä löytyy menetelmällistä kirjallisuutta ja vertaisarvioituja lähteitä. Lisäksi jos toiminnan tavoitteet ovat selkeät, tarkoitus on laskea sanoja ja tehdä konkordanssihakua, on Wordsmith oiva työkalu. Asetuksia voi säädellä melko monipuolisesti ja ne säilyvät omalla koneella tallessa seuraavaa käyttöä varten.

Wordsmith on tavallinen tietokoneohjelma, josta saa pikakuvakkeen työpöydälle. Se on siis luonteeltaan Voyant toolsia aavistuksen yksinkertaisempi käynnistää. Siitä kuitenkin puuttuu erityisesti visuaalisia näkymiä ja

myös Voyant tools on helppokäyttöinen, sekä siihen voi syöttää monipuolisempia erilaisia tiedostomuotoja.

PERUSINFO:
Kehittäjä: Lexical Analysis Software Ltd
Julkaisuvuosi: alkujaan 1996
Kehitysvaihe: uusin versio 8.0 päivitetään edelleen - uusia ominaisuuksia tulee sekä yhteensopivuutta uusiin käyttöjärjestelmiin ylläpidetään.
Hinta: versiot 5.0 ja 4.0 ilmaiset, 50 puntaa kappale
Laitevaatimukset: 5.0: 512 MB keskusmuistia, 40 Mb kovalevytilaa. 8.0: 1GB keskusmuistia, 100 Mb kovalevytilaa.
Käyttöjärjestelmätuki: Windows
Mobiilituki: Ei

AntConc

AntConc on korpusanalyysityökalupakki konkordanssihakuja ja tekstianalyysiä varten. Kuten Wordsmith, myös Antconc on ollut kattavasti käytettynä maailmalla. Antconc on freewarepohjalta pääasiassa yhden tutkijan varassa kehitelty ohjelma. AntConcissa on useampi erilainen työkalu Wordsmith 5.0:an verrattuna.

Kulmikkaampi ja teknisempi

AntConcin lataus ja käyttö on näistä arvosteltavista ohjelmista perinteisin ja selkein omalla koneella: koneelle ladataan asennustiedosto ja sitten ohjelma avataan koneella kuin mikä tahansa muu ohjelma.

Voyant Toolsiin verrattuna AntConc on kulmikkaampi ja teknisempi: sen käytössä ei heti avaudu intuitiivisia ikkunoita, vaan toiminnot ovat usein kielentutkijoiden ammattikielellä nimettyjen valikoiden takana. Käyttäjän pitää tietää etukäteen mikä kukin työkalu on ja mitä se tekee. Lisäksi käyttöjärjestelmässä on avoinna yksi työkalu kerrallaan, mistä suurimpana miinuksena on se, että yksittäistä asiakirjaa tarkistellessa muut näkymät ovat poissa näytöltä.

Ohjelma edellyttää siis vähän perehtymistä ja opettelua. Voyant Tools avaa eri näkymiä heti rinnakkain ja niistä useimmista näkee päälle päin heti mistä on kyse – mikä on visuaalinen apuväline, mikä varsinainen haku ja niin edelleen.

AntConc on hyvin monipuolinen ja sitä voi säätää asetuksista monella eri tavalla itselleen mieleiseksi. Tämä on toki myös heikkous, koska toimintojen monipuolisuus ja ylenmääräinen teknisyys eivät välttämättä ole toimittajalle hyveitä. Toisin kuin Wordsmithin uusimmat versiot, AntConc on täysin

ilmainen (freeware) yksityiseen käyttöön. Sen sijaan yrityksien pitäisi maksaa käyttäjäkohtainen lisenssimaksu. Siksi se saattaa olla hankala käytännössä esimerkiksi mediataloille. Freelance-toimittajalla ei pitäisi olla ongelmaa.

Pidempiin projekteihin

Mikäli toimittajalla on aikaa ja halua perehtyä ohjelmaan ja sen ominaisuuksiin, voi AntConc olla tehokas apuväline. Sitä kun voi räätälöidä mieleisekseen ja säädöt pysyvät tallessa ensi kertaa varten. Projektejaan voi tallentaa ja jatkaa seuraavalla kerralla, eli isommissa tutkivissa hankkeissa tästä voi olla ajan mittaan suurempi hyöty kuin Voyant Toolsista. Sen sijaan perustoimittajan perustyöhön siinä on liian suuri oppimiskynnys. Voyant Tools on nopeampi omaksua ja käyttää.

PERUSINFO:
Kehittäjä: Anthony Laurence
Julkaisuvuosi: 2002
Kehitysvaihe: valmis, päivitetään mikäli yhteensovitusongelmia uusilla käyttöjärjestelmillä
Hinta: Maksuton
Laitevaatimukset: Keskusmuistia käyttää paljon - mitä enemmän aineistoa, sitä kovemmat vaatimukset
Käyttöjärjestelmätuki: Windows, Mac, Linux
Mobiilituki: Ei ole

Overview

Overview on varta vasten journalistista tiedonhankintaa varten tutkimushankkeiden pohjalta luotu tekstinlouhinta- eli haku- ja tekstianalyysiohjelma. Toimittajille räätälöinnin vuoksi ohjelma on kiinnostava. Ohjelman luoneet tutkijat ovat kirjoittaneet myös hyödyllisiä tutkimusartikkeleja ja muita julkaisuja, joissa läpikäydään toimittajien kannalta olennaisia asioita tämänkaltaisille ohjelmille.

Ohjelmassa onkin toimittajan työn kannalta selkein käyttöliittymä: lataat aineiston ja sen jälkeen voit valikoiduilla toiminnoilla analysoida aineistoa. Ohjelman näkökulman on yhteyksien hahmottaminen aineistosta. Muista ohjelmista poiketen tutkijat ovat korostaneet visuaalisia hahmotustyökaluja, eli esimerkiksi vesiputousnäkymää valikoitujen asiasanojen jakautumisesta ja ilmenemisestä eri asiakirjoissa. Tutkijat ovat myös pitäneet tärkeänä toimittajien mahdollisuutta merkitä omia avainsanoja asiakirjoihin ("tag").

Valitettavasti tutkimushankkeet ohjelmasta ovat päättyneet, eikä sitä enää jatkokehitetä. Tämä näkyy erityisesti kahdessa asiassa: asentamisen ja käytön vaikeudessa, ominaisuuksien keskittymisessä englanninkieliseen aineistoon sekä heikoissa näkymätuloksissa. Ohjelma kuitenkin olisi unix-ohjelmointia osaavan räätälöitävissä omia tarpeita paremmin palvelevaksi.

Asentaminen

Overview on ladattavissa ilmaiseksi, joskin ohjelman asentaminen ja käynnistys eivät ole helppoa. Overview:n GitHub -sivuilla on ohjeet muun muassa ohjelman asentamiseksi ja käynnistämiseksi: https://github.com/overview/.

Overview on laadittu Unix-pohjaisena. Toimiakseen Applella tai Windowsilla se vaatii Docker-virtuaalityöskentelytilan. Se taas edellyttää tietokoneen emolevyltä virtuaaliympäristön mahdollistavia toiminnallisuuksia. Käytännössä tämä tarkoittaa, että vain osa uusimmista tietokoneista kykenee Dockeria käyttämään (lisätietoja löydät osoitteesta https://docs.docker.com/desktop/windows/install/).

Paikallinen asennus avautuu selainpohjaisesti. VoyantToolsiin verrattuna ohjelman käyttö sujuu huomattavasti hitaammin ja vaatii enemmän muistia. Asennettu ohjelma tarjoaa helposti vikailmoituksia palvelinasetuksista ja osa toiminnallisuuksista ei ole heti päällä. Mikäli osaa säätää tietokonettaan ja ymmärtää hieman koodin päälle, ovat nämä ongelmat voitettavissa.

Näkymien helppous ja turhuus

Ohjelma käynnistetään lataamalla aineisto ja tämän jälkeen voi valita näkymiä, jotka tarjoavat heti tuloksia muutamalla klikkauksella. Suomenkielisessä aineistossa tulokset vain eivät ole kovin laadukkaita: vaikka ohjelma tunnistaa ruotsin kielen kirjaimet, ei ohjelmaan asennetuista sanakirja- ja korpusaineistoista löydy suomenkielistä vaihtoehtoa.

Englanninkielisiä sanakirjoja ja tietokantoja hyödyntäen näkymissä on esimerkiksi valmiit pikavalinnat paikkakuntien, yritysten tai henkilöiden etsimiseksi aineistosta. Valitettavasti maailmasta löytyy esimerkiksi niin paljon erinimisiä pikkupaikkakuntia, että hutituloksia tulee runsaasti.

Vaikka ohjelman laatineille tutkijoille oli tärkeää, että siihen voi vaivatta syöttää monenlaista aineistoa, oli sitten asiakirjan rakenne tai tiedostotyyppi

millainen vain ("robust import"), niin englanninkielinen testiaineisto sisälsi jonkin verran espanjankielistä tekstiä, mikä sotki useita hakutuloksia. Aineisto oli siis liian "likainen" vierasperäisillä sanoilla.

Esimerkki ikävästä pienestä puutteesta on, että erilaiset analyysilistat rakentuvat siten, että näytetään yleisimmät sanat yleisimmästä alkaen. Nämä listat katkeavat tietyssä vaiheessa, jolloin kaikista harvinaisimmat ja usein mielenkiintoiset yksittäiset sanat jäävät kokonaan näkymättä. Itse haku ohjelmassa vaikuttaisi toimivan varsin hyvin – jos tietää, mitä asiasanaa hakee aineistosta, tarjoaa haku asiakirjat ja kohdat. Lisäämällä tageja ja metadataa asiakirjoihin voi luoda esimerkiksi puunäkymän vain näistä tagatuista asiakirjoista.

Etuna on, että mikäli unix-taitoja löytyy, voi ohjelmaa itse räätälöidä ohjelmoimalla. Tällöin esimerkiksi eri näkymien järjestyksen muuttaminen sekä valmiiden sanakirja-aineistojen tai ohjelman kirjainmerkkituen lisääminen on mahdollista. Mikäli esimerkiksi datajournalistina harrastaa Linux-pohjaista työskentelyä joka tapauksessa, voi koneelleen räätälöidä itselleen Overview-asennuksen suurten data-aineistojen ensivilkaisua varten.

PERUSINFO:
Kehittäjä: The Associated Press news agency - Jonathan Stray
Julkaisuvuosi: 2011 prototyyppi, selainpohjainen uudelleenjulkaisu 2013
Kehitysvaihe: Päättynyt.
Hinta: Ilmainen
Laitevaatimukset: Unix-pohjainen kone ei välttämättä vaadi paljon, mutta muilla käyttöjärjestelmillä edellytyksenä on emolevyn virtuaaliympäristötuki (Dockers:a varten) ja paljon muistia (käytännössä melko uusi kone)
Käyttöjärjestelmätuki: Unix-pohjainen, vattii Dockers-ympäristön Windows/Apple -koneilla
Mobiilituki: Ei ole

Lähteet

Ampofo, L., Collister, S., O'Loughlin, B., & Chadwick, A. (2015). Text Mining and Social Media: When Quantitative Meets Qualitative and Software Meets People. In P. Halfpenny & R. Procter, *Innovations in Digital Research Methods* (pp. 161–192). SAGE Publications Ltd. https://doi.org/10.4135/9781473920651.n8

Attfield, S. & Dowell, J. (2003). Information seeking and use by newspaper journalists. *Journal of Documentation*, 59(2), 187–204. https://doi.org/10.1108/00220410310463860

Bednarek, M., & Carr, G. (2020). Computer-assisted digital text analysis for journalism and communications research: Introducing corpus linguistic techniques that do not require programming. *Media International Australia*. https://doi.org/10.1177/1329878X20947124

Brehmer, M., Ingram, S., Stray, J., & Munzner, T. (2014). Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool for Investigative Journalists. *IEEE Transactions on Visualization and Computer Graphics*, 20(12), 2271–2280. https://doi.org/10.1109/TVCG.2014.2346431

Campbell, F. (1997), "Journalistic construction of news: information gathering", *New Library World*, Vol. 98 No. 2, pp. 60-64. https://doi.org/10.1108/03074809710159330

De Grove, F., Boghe, K., & De Marez, L. (2020). (What) Can Journalism Studies Learn from Supervised Machine Learning? *Journalism Studies*, 21(7), 912-927. https://doi.org/10.1080/1461670X.2020.1743737

Deuze, M. (2005). What is journalism?: Professional identity and ideology of journalists reconsidered. *Journalism*, 6(4), 442–464. https://doi.org/10.1177/1464884905056815

Harjumaa, M. & Jansson, K. (31.8.2020). Bangkok, Kapkaupunki, New York – Virkamatkojen taloudellisuutta valvovan viraston johtaja matkusti viime vuonna 55 000 eurolla. *YLE uutiset*. https://yle.fi/uutiset/3-11514399

Hase, V., Engelke, K. M., & Kieslich, K. (2020). The things we fear. combining automated and manual content analysis to uncover themes, topics and threats in fear-related news. *Journalism Studies*, 21(10), 1384-1402. https://doi.org/10.1080/1461670X.2020.1753092

Heyl, A., Joubert, M., & Guenther, L. (2020). Churnalism and hype in science communication: Comparing university press releases and journalistic articles in south africa. *Communicatio*. https://doi.org/10.1080/02500167.2020.1789184

Hopeakunnas, M. (2015). Katsaus toimittajien käyttämiin lähteisiin ja tietokäyttäytymiseen. *Informaatiotutkimus*, 34(1-2). Noudettu osoitteesta https://journal.fi/inf/article/view/53745

Houston, B., Investigative Reporters & Eds. (2008). Investigative Reporter's Handbook: A Guide to Documents, Databases, and Techniques. Bedford/St. Martin's. Viides painos.

Jyväskylän yliopisto, Kieli- ja Viestintätieteiden laitos. Julkisuuslaki.fi. Julkisuuslakia ja -periaatetta analysoivat Jyväskylän yliopiston tutkimushankkeet. https://julkisuuslaki.fi

Liu, Y., Barlowe, S., Feng, Y., Yang, J., & Jiang, M. (2013). Evaluating exploratory visualization systems: A user study on how clustering-based visualization systems support information seeking from large document collections. *Information Visualization*, 12(1), 25–43. https://doi.org/10.1177/1473871612459995

Paananen, K. & Liski, J. (2.6.2020). Kallista konsultointia. Suomen Kuvalehti, politiikka.

Potts, A., Bednarek, M., & Caple, H. (2015). How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina. *Discourse & Communication*, 9(2), 149–172. https://doi.org/10.1177/1750481314568548

Pöyhtäri, R., Väliverronen, J., & Ahva, L. (2016). Suomalaisen journalistin itseymmärrys muutosten keskellä. *Media & Viestintä*, 39(1). https://doi.org/10.23983/mv.61434

Rämö, A. & Liski, J. (19.2.2018). Saariston Sampo. Suomen Kuvalehti, kotimaa.

Su, Y., Hu, J., & Lee, D. K. L. (2020). Delineating the transnational network agenda-setting model of mainstream newspapers and twitter: A machine-learning approach. *Journalism Studies*. https://doi.org/10.1080/1461670X.2020.1812421

Kirjoittajat

Markus Mykkänen, tutkijatohtori, Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto, ORCID ID: 0000–0002–7044–9263

Aleksi Koski, projektitutkija, Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto, ORCID ID: 0000-0003-0351-4982