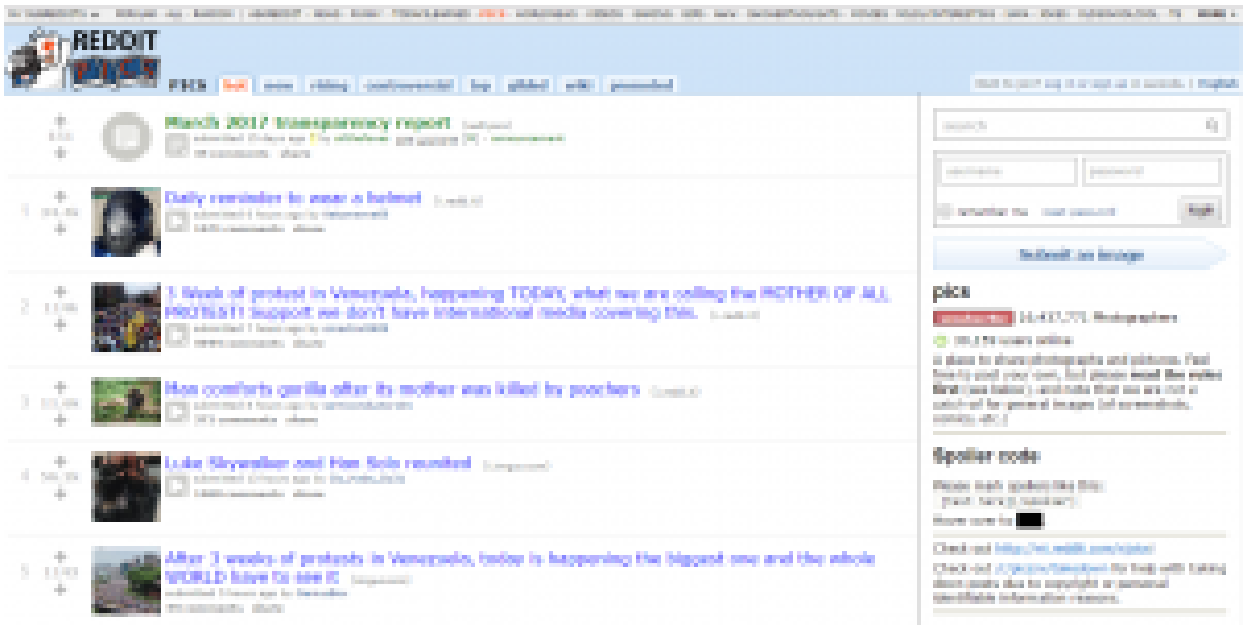
Mitä tilastollinen tarkastelu voi kertoa sosiaalisen median kielestä?

Aatu Liimatta

Reddit on pääasiassa englanninkielinen sosiaalisen median sivusto, jossa keskustelu keskittyy eri aihealueiden ympärille. Tilastollisia menetelmiä käyttämällä voidaan saada selville, että Redditin sisällä kielenkäyttö vaihtelee tilanteen mukaan samoin kuin sen ulkopuolellakin. Ihmiset ovat taitavia käyttämään kieltä aina tilanteen vaatimalla tavalla.

Mikä Reddit?

Sosiaalisen median sivusto Reddit (www.reddit.com) on suhteellisen tuntematon, vaikka se on kolmanneksi suosituin englanninkielinen sosiaalinen media Facebookin ja Twitterin jälkeen. Reddit koostuu niin kutsutuista *aliredditeistä* (*subreddit*) eli eri aihealueita käsittelevistä alifoorumeista. Kuka tahansa voi aloittaa uuden alireddin, joten Redditin aihealueet kattavat kaiken kuviteltavissa olevan maan ja taivaan väliltä (ja paljon sellaista mitä ei voi kuvitellakaan). Redditin käyttäjät voivat tehdä postauksia eri aliredditeihin. He voivat myös kommentoida muiden tekemiä postauksia ja vastata kommentteihin. Näin käydyt keskustelut ovat varmastikin Redditin suosituinta sisältöä. Käyttäjät voivat myös "tilata" haluamansa alireddit, jolloin niiden viestiketjut näkyvät heidän Reddit-etusivullaan.



Kuva 1. Mielenkiintoisille kuville tarkoitettu alireddit, /r/Pics,.

Kuva 2. Keskustelua kypärien tärkeydestä ja rekkojen turvallisuudesta /r/Pics-alireddittistä.

Alireddition aiheet voivat olla niinkin yleisiä kuin uutiset (/r/worldnews), politiikka (/r/politics), kuvat (/r/Pics), televisio (/r/television), pelit (/r/games), vitsit (/r/jokes), matematiikka (/r/mathematics) tai kieli (/r/language), tai hyvinkin tarkkaan rajattuja,

kuten tiettyjä kaupunkeja, tiettyjä poliittisia näkökulmia tai tiettyjä elokuvia, televisiosarjoja ja tietokonepelejä käsittelevät alireddit. Myös vaikkapa monille sisäpiirin vitseille on omat alireddit. Monet alireddit ovat aiheeltaan lähellä toisiaan, mutta käsittelevät aihetta hiukan eri näkökulmista tai omalla tavallaan. Esimerkiksi "historia"-alireddit (/r/History) on tarkoitettu kaikelle keskustelulle ja linkeille historiaan liittyen, kun taas "kysy historiantutkijoilta" (/r/AskHistorians) on tarkkaan moderoitu alireddit, jossa historiantutkijat ja asiansa osaavat harrastajat vastaavat esitettyihin historiaan liittyviin kysymyksiin lähes akateemisin standardein.

Tilastollista rekisteritutkimusta

Mutta mitä tilastollinen tarkastelu voi kertoa Redditin kielestä? Paljonkin, riippuen siitä, mitä halutaan tarkastella. Itse olen kiinnostunut rekisteritutkimuksesta, joka vertailee eri tilanteiden kielenkäytölle tyypillisiä ja epätyypillisiä piirteitä. Lehtiuutinen kirjoitetaan eri sanankäänteinkuin postikortti lomamatkalta, ja poliitikko puhuu toimittajille eri tavoin kuin opettaja luokalleen. Mutta miten on Redditin laita? Miten eri alireddittien kielenkäyttö eroaa toisistaan? Miten tällaisia eroja voidaan tutkia? Yksittäisten tekstiesimerkkien vertailu, kuten yhden postikortin vertaaminen yhteen uutisartikkeliin, tai yhden Reddit-viestin vertaaminen toiseen, saattaa kertoa jotakin näiden kielimuotojen eroista, mutta todennäköisesti kertoo enemmän postikortin tai artikkelin kirjoittajan yksilöllisestä kielenkäytöstä kuin postikorttien tai uutisartikkelien kielenkäytöstä yleensä. Niinpä on tarpeen ottaa käyttöön suurten tekstiaineistojen eli korpusten tilastollinen vertailu.

Yksi tunnetuimpia ja käytetyimpiä rekisteritutkimuksen korpusmenetelmiä on Douglas Biberin jo 1980-luvun loppupuolella kehittämä "moniulotteinen rekisterianalyysi". Tätä menetelmää käytettäessä vertaillaan useita kymmeniä kielen peruspiirteitä kuten eri sanaluokkia, verbien aikamuotoja, eri persoonapronomineja ja erilaisia rakenteita. Näiden rakenteiden tekstikohtaiset esiintymistiheydet lasketaan tietokoneohjelman avulla. Eri piirteet esiintyvät eri teksteissä eri tiheyksillä. Esiintymistiheyksiä tilastollisesti vertailemalla voidaan löytää piirrekimppuja, joiden piirteet tapaavat esiintyä teksteissä yhdessä ja vastaavasti olla poissa teksteistä yhtä aikaa. Jokainen tällainen kimppu muodostaa "rekisteriulottuvuuden": onhan olemassa jokin syy, miksi nämä piirteet esiintyvät yhdessä ja miksi kirjoittaja tai puhuja on päättänyt käyttää (tai olla käyttämättä) näitä piirteitä.

Tällä menetelmällä on pitkät perinteet. Jo vuonna 1988 Biber vertaili laajasti 23 englannin kielen puhuttua ja kirjoitettua tekstilajia, mm. erilaisia faktatekstejä kuten

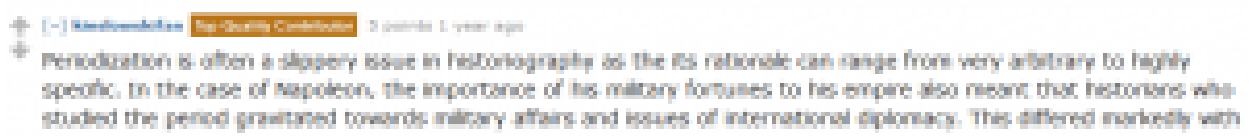
uutisjuttuja, pääkirjoituksia, elämäkertoja ja virallisia asiakirjoja; eri fiktiogenrejä; ja puhuttua kieltä kuten keskusteluja, haastatteluja ja puheita. Hän löysi näistä tekstilajeista kuusi rekisteriulottuvuutta. Näistä tärkein ja tunnetuin on ensimmäinen ulottuvuus, "osallistuva tai informatiivinen tuotto" (Involved vs. Informational Production). Toisin sanottuna kaikki tutkitut tekstit ja tekstilajit asettuvat jollekin kohtaa akselia, jonka toisessa päässä sijaitsevat osallistuvat tekstilajit, Biberin materiaalissa ennen kaikkea puhelinkeskustelut, ja toisessa päässä tiiviisti tietoa sisältävät tekstilajit kuten akateeminen teksti ja uutistekstit. Osallistuville teksteille tyypillisiä piirteitä englannin kielessä Biberin tutkimuksen mukaan ovat esimerkiksi ihmisten ajattelua ja mielipiteitä ilmaisevat verbit ja preesensmuotoiset verbit ylipäänsä sekä lyhennetyt muodot kuten *I'm* ja *can't* (eikä *I am* ja *cannot*) ja toisen persoonan pronomini *you*. Informatiivisissa teksteissä nämä piirteet taas ovat harvinaisempia. Sen sijaan informatiiviset tekstit sisältävät paljon substantiiveja, ja niiden sanat ovat keskimäärin pidempiä ja vaihtelevampia. (Biber 1988.)

Redditin rekisteriulottuvuudet

Teen väitöskirjaa tilastollisten menetelmien soveltamisesta sosiaalisen median ja erityisesti Redditin rekisteritutkimukseen, ja alustavien havaintojen mukaan myös Redditin sisältä voi Biberin menetelmän avulla löytää selkeitä rekisteriulottuvuuksia. Tarkastelemalla 27 alirediditin joukkoa (johon kuuluu mm. aiemmin mainitsemani /r/AskHistorians) olen saanut selville, että analysoimani alireddit asettuvat ainakin kolmelle rekisteriulottuvuudelle. Tietenkin on pidettävä mielessä, että kaikki alireddit sijoittuvat joka ulottuvuudella jonnekin kahden ääripään välille, ja monissa alirediteissä ulottuvuuden kaksi napaa ovat varsin hyvin tasapainossa.

Ensimmäinen ulottuvuus, joka selittää suurimman osan rekisterivaihtelusta alireddittien välillä, on "henkilöfokus tai asiafokus" (Personal vs. Factual Focus). Jotkin alireddit, kuten japanilaisia tarinapelejä käsittelevä /r/visualnovels, keskittyvät voimakkaammin henkilöihin, ajatuksiin ja mielipiteisiin; toiset taas keskittyvät faktoihin ja asiatietoon, kuten /r/AskHistorians tai tietokoneeseen kytkettäviä radiovastaanottimia käsittelevä /r/RTLSDR. Ulottuvuuden asiafokus-pään alireddit eille tyypillinen kielenpiirre on substantiivien suuri määrä. Henkilöfokus-pään alirediteissä käytetään vähemmän substantiiveja, mutta sen sijaan runsaammin henkilön ajattelua kuvaavia verbejä kuten *think*, *feel*, *understand*, *assume* (ajatella, tuntea, ymmärtää, olettaa) ja sanallista ulosantia kuvaavia verbejä kuten *say, admit, suggest, claim* (sanoa, myöntää, ehdottaa, väittää) sekä mm. adverbejä kuten *quickly* (nopeasti) ja lyhennettyjä muotoja kuten *I'm* ja *can't*.

Toinen ulottuvuus on "informatiivinen tai osallistuva tyyli" (Informational vs. Involved Style). Kuten yllä mainitsemani Biberin ensimmäinen ulottuvuus, tämä Redditin ulottuvuus kuvastaa sitä, kuinka keskusteleva alirediditin tyyli on, vai sisältääkö se enemmän tiiviin informatiivisia tekstejä. Toiset alirediditit ovat tyyliltään enemmän keskustelevia, kuten kauniisiin luontokuviin keskittyvä /r/EarthPorn; toiset, esimerkiksi /r/AskHistorians, taas sisältävät paljon informatiivista, tiivistä tekstiä. Informatiivisen pään teksteille tyypillisiä piirteitä ovat mm. keskimäärin pidemmät sanat ja nominalisaatiot eli muista sanaluokista muodostetut substantiivit kuten *movement* (liike, verbistä *move*) tai *carelessness* (huolimattomuus, adjektiivista *careless*). Osallistuvalle tyylille näiden piirteiden sijaan tyypillisiä ovat ennen kaikkea ensimmäisen persoonan pronominit kuten *I* ja *me*.


[–] Periodization is often a slippery issue in historiography as the its rationale can range from very arbitrary to highly specific. In the case of Napoleon, the importance of his military fortunes to his empire also meant that historians who studied the period gravitated towards military affairs and issues of international diplomacy. This differed markedly with

Kuva 3. /r/AskHistorians on tyypillinen esimerkki asiafokuksen ja informatiivisen tyylin alirediditistä. Huomaa mm. sanojen pituus, substantiivien määrä, ja nominalisaatio *periodization*, sekä ensimmäisen persoonan pronominien ja lyhennettyjen muotojen puute.


[–] I don't know the actual name of it, but there is a kind of vegetable that looks exactly like corn, but it's miniature. It tastes nothing like corn, and isn't actually corn at all. I think it was a kind of squash...?

Kuva 4. /r/whatisthisthing-alirediditissä käyttäjät auttavat toisiaan tunnistamaan tuntemattomia asioita.

Tämä esimerkki on rekisteriltään päinvastainen kuin ylempi /r/AskHistorians-esimerkki: sanat ovat lyhyempiä ja substantiiveja on vähemmän, mutta sen sijaan ajattelua kuvaavia verbejä kuten *think* ja *know*, lyhennettyjä muotoja ja ensimmäisen persoonan pronomineja on enemmän.

Kolmas ulottuvuus on "nykyajan tai menneen ajan fokus" (Non-Past vs. Past Focus). Jotkin alirediditit, mukaan lukien /r/AskHistorians, keskittyvät menneeseen aikaan, esimerkiksi menneisiin tapahtumiin, historiaan tai tarinoiden kertomiseen, kun taas toiset, kuten algoritmeja käsittelevä /r/algorithms, keskittyvät voimakkaasti nykyaikaan

tai abstrakteihin asioihin. Tämän ulottuvuuden ääripäitä luonnollisesti hallitsevat eri aikamuodot, nykyajan teksteissä preesens ja menneen ajan teksteissä mennyt aikamuoto, mutta nykyajan teksteihin liittyy myös paljon muita piirteitä kuten erilaisia tulevaisuutta ja tulevaisuuden mahdollisuuksia kuvaavia apuverbejä (esim. *will*, *can*, *may*) ja alistuskonjunktiot *if* ja *unless*.

Kieli on kommunikaation väline

Mutta eivätkö nämä ulottuvuudet, "henkilöfokus tai asiafokus", "informatiivinen tai osallistuva tyyli" ja " nykyajan tai menneen ajan fokus", kuulosta siltä, että minkä tahansa tekstin maailmassa pystyisi luokittelemaan niiden avulla, eikä vain Redditin alireddite jä? Kyllä vain, ja mielestäni nimenomaan tässä piilee asian kauneus. Kielestä, kielimuodosta tai kielenkäytön tilanteesta riippumatta moniulotteinen rekisterianalyysi tapaa löytää aina tiettyjä samantyyppisiä rekisteriulottuvuuksia, mutta myös tutkimuksen rajauksen mukaan uniikkeja, nimenomaan kyseiseen kielen muotoon tai tilanteeseen liittyviä ulottuvuuksia (Biber 2014; vrt. Biber 1995, Biber 2016).

Reddit ei siis tässä suhteessa eroa muista ihmisten käyttämistä kommunikaatiokanavista. Paljolti samantyyppiset tilannekohtaiset vaatimukset asettavat rajoja kielelle niin Redditissä kuin sen ulkopuolellakin. Mutta toisaalta jokaisen kommunikaatiokanavan vaatimukset ovat erilaiset, ja siten joka tilanteen kielelliset ratkaisut ovat uniikkeja.

Loppujen lopuksihan kieli on kommunikaation väline. Usein haluamme luokitella tekstejä ja kieltä eri tavoin: puhuttua tai kirjoitettua kieltä, nuorten tai aikuisten kieltä, uutisten tai kaunokirjallisuuden kieltä, hyvää tai huonoa kieltä. Olen ehkä puolueellinen, mutta väitän, että rekisteritutkimus on yksi parhaita tapoja nähdä, että kieli ei ole sellainen tarkasti lokeroitava, jäykkä järjestelmä tiukoilla säännöillä kuin usein helposti kuvittelemme, vaan oikeastaan ääretön määrä mahdollisuuksia, joita kielen käyttäjät osaavat soveltaa ja hyödyntää aina tilanteen vaatimusten ja tarpeiden mukaan, jotta tehokas ja asianmukainen kommunikaatio säilyisi, niin sosiaalisessa mediassa kuin sen ulkopuolellakin. Ei postikorttiakaan kirjoiteta kuin lehtiuutista.

Kirjoittaja on englantilaisen filologian tohtorikoulutettava Helsingin yliopiston nykykielten laitoksella.

Lähteet

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in contrast, 14*(1), 7–34.

Biber, D., & Egbert, J. (2016). Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics, 44*(2), 95–137.