Mallivastaukset

Tehtävä 1: Historiallinen korpuslingvistiikka

1. Nevalainen ym. (2020) käsittelevät makrotason laajempaa kysymystä kielenmuutoksen nopeuteen vaikuttavista ulkoisista tekijöistä, Budts ja Petré (2016) puolestaan mikrotason kysymystä yksittäisen kielen ilmiön eli *be going to* -rakenteen kehityksestä.
2. Artikkelissa käsitellään aineiston tasapainon muutosta ajassa, joka on saattanut osaltaan vaikuttaa tutkimuksessa havaittuun sukupuolierojen pienenemiseen. On vaikeaa arvioida, missä määrin tuloksiin vaikuttaa aineiston epätasapaino ja missä määrin kyse on oikeasta kielenkäytön muutoksesta, johon vaikuttivat yhteiskunnalliset tekijät.

Tehtävä 2: Historialliset korpukset

1. EMEMT-korpus on yhden genren eli lääketieteellisten tekstien korpus. Se on diakroninen, eli siihen on koottu tekstejä useilta eri aikakausilta vuosien 1500 ja 1700 väliltä, ja sitä voidaan käyttää ajallisen muutoksen tutkimiseen. Sitä voidaan luonnehtia rikkaaksi dataksi, koska korpusteksteistä on saatavilla metadatana kirjoitusvuoden lisäksi tarkat tekstikategoriatiedot. Isona datana en tuollaista kahden miljoonan sanan korpusta pitäisi, koska se on vielä periaatteessa ihmisluettavan kokoinen, siitä saadut hakutulokset voidaan käydä tarkasti läpi kohtuullisessa ajassa ja koko on suhteellisen pieni verrattuna megakorpuksiin ja moniin tekstitietokantoihin.
2. Stefanowitschin mukaan edustavuutta on ensinnäkin vaikea arvioida, koska meillä ei yleensä ole tietoa siitä, miten relevantit parametrit (kuten se, miten paljon eri tekstilajeja tuotetaan) ovat jakautuneet populaatiossa, jota pyritään edustamaan (esim. tietyn kielen puhujat jonakin aikakautena). Myöskään eri tekstilajien vastaanottajien määristä ei ole tarkkaa tietoa; vaikka jotakin tekstilajia olisi tuotettu hyvin paljon, sen lukijoiden määrä voi olla hyvinkin pieni, joten on epäselvää, miten näitä pitäisi painottaa. Kieliyhteisöt eivät myöskään ole mitenkään homogeenisiä, joten mikään valittu tasapaino parametrien välillä ei voi edustaa kaikkia

kielenpuhujia. On myös se käytännön ongelma, että joistakin genreistä, kuten vaikkapa luottamuksellisista keskusteluista asianajajan ja asiakkaan välillä, ei pystytä saamaan otosta korpukseen. Stefanowitsch ehdottaakin, että edustavuuden ja tasapainon sijaan puhuttaisiin monipuolisuudesta eli diversiteetistä: korpuksen kokoajat pyrkivät käytännössä edustamaan valitsemaansa kielimuotoa mahdollisimman monipuolisesti valitsemalla tekstejä mahdollisimman monesta eri varieteetista.

- Stefanowitschin mukaan Brown-korpus edustaa korkeintaan niiden kirjastojen kokoelmia, joiden perusteella se on koottu, joskaan tapa, jolla tekstejä on eri kategorioista valittu, ei ole erityisen edustava, ja korpukseen on lisätty aineistoa myös kokoelmien ulkopuolelta.
- Kielten varhaisempaa historiaa tutkittaessa ongelmaksi muodostuu mm. se, että puheaineistoa ei ole saatavilla, jolloin korpus ei voi edustaa normaalien arkikeskusteluiden kieltä, jota ihmiset oikeasti kohtaavat ja tuottavat kaikkein eniten. Läheskään kaikki ihmiset eivät olleet kirjoitustaitoisia, joten tekstiaineistot edustavat usein vain hyväosaisten miesten kieltä. Lisäksi tekstilajien välinen tasapaino on hankala säilyttää ajassa, koska genret muuttuvat, niitä tulee uusia ja jotkin häviävät. Englannin kieli on siinä mielessä onnellisessa asemassa, että tekstiaineistoa on tuotettu ja säilynyt kielen varhaisvaiheista nykypäivään asti, joskin muinaisenglannista säilynyt aineisto on hyvin rajallinen, ja keskienglantiin siirryttäessä tekstit tulevat eri murrealueelta, mikä vaikeuttaa kielenmuutoksen tutkimista.

Tehtävä 3a: Aineiston hakeminen AntConcilla

1. Etsin *hath*-muotoa näillä hakusanoilla, joita voit verrata omiisi: *hath*, *hathe*, *hatht*, *heth*, *hethe*.
2. Tässä kaksi mahdollista hakulauseketta selityksineen:
   - **\b[Hh](a(['y]?se?|ce)|ese?)\b**
     - **\b** = sanaraja, sitten tulee joko iso tai pikku H **[Hh]**, sitten tulee joko a:llisia tai e:llisiä muotoja. Nämä kaksi vaihtoehtoa, a:lliset muodot eli **a(['y]?se?|ce)** ja e:lliset muodot eli **ese?**, ovat lausekkeessa sulkeiden **()** sisällä ja putkimerkillä **|** eroteltuina **(a(['y]?se?|ce)|ese?)**.

        - **a**-muotojen sisällä on kaksi vaihtoehtoa, joissa sibilantti on kirjoitettu joko s:llä tai c:llä. Nämä kaksi vaihtoehtoa ovat jälleen lausekkeessa sulkeiden sisällä ja putkimerkillä eroteltuina **(['y]?se?|ce)**.
            - s-vaihtoehdossa a:n jälkeen tulee mahdollisesti joko heittomerkki tai y **['y]?**, sitten tulee **s**, ja lopuksi tulee mahdollisesti **e?**. Tämä haku osuu siis muotoihin *has*, *hase*, *ha's*, *ha'se*, *hays* ja *hayse* (isolla tai pienellä alkukirjaimella).
            - c-vaihtoehdossa a:n jälkeen tulee **c** ja **e**. Tämä osuu muotoon *hace*.
        - **e**-muodoissa tulee e:n jälkeen **s** ja loppuun mahdollisesti **e?**. Tämä osuu muotoihin *hes* ja *hese*.
    - Lopuksi sana päättyy **\b**. Tämän haun tulokset vastaavat täysin sanalistalla tekemäämme hakua.
- **\b[Hh][ae]y?'?[sc]e?\b**
    - **\b** = sanaraja, sitten tulee joko iso tai pikku H **[Hh]**, sitten tulee joko a tai e **[ae]**, sitten tulee mahdollisesti **y?**, sitten tulee mahdollisesti heittomerkki **'?**, sitten tulee joko s tai c **[sc]**, sitten tulee mahdollisesti **e?** ja lopuksi sana päättyy **\b**.
    - Tämä haku ottaa laajemmin eri vaihtoehtoja huomioon. Tee haku AntConcissa ja katso, mitä saat tulokseksi!

Tehtävä 3b: Aineiston käsittely Excelissä

- Voit verrata tiedostoasi soveltuvin osin tehtävän 3d vastauksessa annettuun Excel-tiedostoon.

Tehtävä 3c: Normalisoidun frekvenssin laskeminen ja visualisointi

1. Voit verrata tiedostoasi soveltuvin osin tehtävän 3d vastauksessa annettuun Excel-tiedostoon.
2. Voit verrata tuloksiasi tästä osiosta löytyvään Excel-tiedostoon ”hashath-variaabeli.xlsx”, jossa *has*- ja *hath*-muotoja on käsitelty variaabelina.

Tehtävä 3d: Tilastollinen merkitsevyys

1. Voit verrata tuloksiasi tästä osiosta löytyvään Excel-tiedostoon ”hath.xlsx”, jossa on analysoitu *hath*-muotoa ja johon on myös kopioitu GraphPadin antamat merkitsevyystulokset.
2. Tekstien lukumäärään perustuva merkitsevyystestaus toimii parhaiten silloin, kun tutkittavan ilmiön esiintymistiheys on keskitasoa (ks. tarkemmin alla). Sen etuna on dispersiotietoisuus sekä suhteellinen vaivattomuus (vrt. Vartiainen & Säily 2020: 540–541, ks. osion kirjallisuusluetteloa).
    - Aineistosta pitää tietää tai pystyä laskemaan (a) tekstien lukumäärä kullakin aikakaudella ja (b) niiden tekstien määrä, joissa tutkittua ilmiötä esiintyy kullakin aikakaudella. Näiden tietojen perusteella pystyy myös laskemaan niiden tekstien määrän, joissa ilmiötä ei esiinny. Joistakin korpuksista tietoa tekstien lukumäärästä per aikakausi ei ole helposti saatavilla, esimerkiksi Mark Daviesin kokoamista englannin, espanjan ja portugalin kielen megakorpuksista, jotka ovat hänen suunnittelemansa web-käyttöliittymän takana.
    - Jos ilmiö on hyvin harvinainen, se saattaa esiintyä vain muutamassa tekstissä, jolloin nollahypoteesia ei pystytä kumoamaan, koska pienissä määrissä esiintyvät muutokset ovat väkisinkin pieniä, jolloin ne voivat olla pelkkää sattumaa. Sama ongelma on myös sellaisissa merkitsevyystesteissä, joissa vertaillaan ilmiön esiintymistiheyttä tekstitason sijaan sanatasolla (moniko sana edustaa vs. ei edusta ilmiötä kullakin aikakaudella), mutta se tulee vastaan vähän vähemmän herkästi.
    - Jos ilmiö on hyvin yleinen (esim. englannin määräinen artikkeli), se saattaa esiintyä jokaisessa tekstissä, jolloin nollahypoteesia siitä, että aikakausien välillä ei ole eroa ilmiön käytössä, ei taaskaan pystytä kumoamaan. Tällöin toimivat paremmin sellaiset merkitsevyystestit, joissa vertaillaan ilmiön esiintymistiheyttä sanatasolla: vaikka määräinen artikkeli esiintyisikin jokaisessa tekstissä, sen frekvenssi tekstien sisällä voi toki muuttua ajassa.
    - Jos ilmiötä käsitellään variaabelina, ei kannata laskea tekstejä, koska samassa tekstissä voi esiintyä kumpaakin varianttia. Tällöinkin on mentävä sanatasolle ja laskettava, kuinka moni sana edustaa toista varianttia ja kuinka moni toista.

Tehtävä 3e: Muutoksen selittäminen

1. Eniten *hath*-muotoa löytyy saarnoista, ja uskonnollisissa teksteissähän *hath*-muotoon saattaa törmätä vielä nykyäänkin.
    - *Hath*-muotoa on yllättävän paljon yksityiskirjeissä, joihin *has*-muoto saapui jo varhain. Ehkäpä yksityiskirjeet ovat muuttuneet konservatiivisemmiksi varhaisuusenglannin kauden loppua kohti.
2. Tuloksien tulkinnassa voidaan käyttää apuna mm. aiempaa tutkimusta aiheesta, aikakauteen liittyvää historiantutkimusta sekä aikalaiskirjoituksia, joista löytyy metalingvistisiä kommentteja.
3. En keksi mitään muuta menetelmää, jolla muutoksesta voisi saada tällaisen kvantitatiivisen yleiskuvan.
    - Koska aineistomme on pelkkää kirjakieltä, emme voi saada tarkkaa kuvaa siitä, miten muutos on puhekielessä edennyt, joskin vaikuttaa vahvasti siltä, että muutos on alkanut puhekielestä ja näkyy teksteissä viiveellä. Koska suurin osa kielenkäytöstä tapahtuu puheen välityksellä ja kirjoitustaitoisten ihmisten osuus on eri aikoina voinut olla hyvinkin pieni, puheen tutkiminen olisi tärkeää mutta historiallisessa kielentutkimuksessa usein mahdotonta. On kuitenkin joitakin tekstilajeja, jotka ovat lähellä puhekieltä, kuten yksityiskirjeet, ja niitä ovat kirjurien välityksellä luoneet myös kirjoitustaidottomat. Lisäksi kirjurit ovat taltioineet puhetta mm. oikeudenkäyntipöytäkirjoihin.
    - Aineiston niukkuus on saattanut vaikuttaa esim. siihen, ettemme pystyneet osoittamaan, että kahden ensimmäisen aikakauden välinen ero olisi tilastollisesti merkitsevä. Ero tosin on selvästi pienempi kuin kahden viimeisen aikakauden välinen ero, joten aineiston määrän kasvattaminen ei välttämättä olisi auttanut asiaa. Mitä enemmän aineistoa on, sitä paremmin tuloksiin voi luottaa, ellei iso aineisto ole epätasapainoisempi kuin pieni.
    - Aineiston genretasapaino muuttuu hieman ajassa, kuten tehtävän yhteydessä esitetystä kuvasta näkyy. Esimerkiksi uskonnollisten tekstien osuus pienenee viimeisellä aikakaudella. Koska tiedämme, että *hath*-muotoa esiintyy etenkin uskonnollisissa teksteissä, voi osa sen näennäisestä häviämisestä selittyä sillä, että genreä, jossa sitä erityisesti esiintyy, on korpuksessa vähemmän. Tätä kysymystä voitaisiin selvittää laskemalla

normalisoituja frekvenssejä genreittäin: jos *hath*-muodon frekvenssi pienenee ajan myötä useimmissa genreissä, ei muutos voi johtua pelkästään siitä, että uskonnollisten tekstien määrä korpuksessa vähenee. Uskonnollisten tekstien osuus ei myöskään ole millään aikakaudella niin suuri, että koko havaittu kielenmuutos voisi selittyä pelkästään sen muutoksella. Vaikka aineiston epätasaisuus on ehkä tutkimuksen merkittävin ongelma, se ei kuitenkaan ole mielestäni tässä tapauksessa erityisen suuri ongelma, ja tutkimustuloksia voidaan pitää uskottavina.