Sini Knuutila, Olli Kuparinen, Jenni Santaharju,
Liisa Mustanoja, Unni Leino ja Jaakko Peltonen

Miksi kato leviää?

hd-*yhtymän katovariantin diffuusion syyt Helsingin puhekielessä*

1 Johdanto

Tässä artikkelissa tutkimme yleiskielen *hd*-yhtymän katovariantin esiintymisen syitä. Aiemmin on havaittu (Kuparinen 2021), että *hd*-yhtymien soinnillisen klusiilin kato on ekspansiivinen ilmiö, kun tarkasteltavana on lähes viisi vuosikymmentä kattava kolmen aikapisteen aineisto Helsingissä puhuttua suomea. Aiemmassa tutkimuksessa ei ole kuitenkaan juuri otettu kantaa siihen, miksi katoa tapahtuu ja mitkä tekijät siihen tarkalleen ottaen vaikuttavat. Muiden ilmiöiden osalta on arveltu, että mahdollisesti sanan suuri frekvenssi olisi erityisen merkityksellinen tekijä muutoksen etenemisessä (ks. esim. Lehtimäki 1983; Suihkonen 1992; Bybee 2002). Frekvenssi vaikuttanee myös katomuotojen esiintymiseen, sillä eniten katomuodoissa esiintyvät lekseemit ovat hyvin taajaan esiintyviä (Kuparinen–Santaharju–Leino–Mustanoja–Peltonen 2022). Niin ikään on arveltu, että eräät seikat, kuten *hd*-yhtymän sijoittuminen konsonanttivartaloon (*tehdä*) ja esiintyminen painottoman tavun (*pyörähdä*-) tai pitkän vokaalin (*hiihdä*-) jäljessä, saattavat hidastaa katomuotojen etenemistä (mts. 32). Tommi Kurki (2005, 122–123) on osoittanut, että jos lekseemi kuuluu astevaihtelun piiriin, kadon todennäköisyys pienenee.

Kuparinen kollegoineen (2022) osoitti, miten kato etenee leksikaalisesti ja morfologisesti. Tässä artikkelissa tarkastelemme, mitkä kielensisäiset syyt vaikuttavat kadon leviämiseen. Tutkimme erityisesti, miten katovarianttia kantavan lekseemin frekvenssi ja merkitys sekä *hd*-yhtymän sijainti sanan sisällä (esim. vartalotyyppi, painoasema, astevaihteluasema ja edeltävän vokaalin laatu) vaikuttavat kadon todennäköisyyteen. Kadon taustalla vaikuttavien tekijöiden tunteminen on tärkeää, jotta kadon etenemistä ja sitä myöten kielen kehityksen suuntaa voidaan ennustaa. Lähtökohdiltaan nykyinen tutkimus on samansuuntainen kuin Joan Bybeen (2002) tutkimus amerikanenglannin *t*:n ja *d*:n sananloppuisesta kadosta: Bybee havaitsi sanan korkean frekvenssin edistävän katoa, mutta esiintymiskontekstin vaikuttavan joko hidastavasti tai nopeuttavasti katomuotojen diffuusioon. Menetelmällisesti tutkimukset kuitenkin eroavat siinä, että tässä tutkimuksessa kielensisäisten muuttujien yhdistelmiä on tarkasteltu logistisen regression ja päätöspuumallin avulla.

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Artikkeli rakentuu niin, että seuraavassa luvussa esitellään artikkelissa käytetyt avaintermit ja aiempi aihetta koskeva tutkimus. Kolmannessa luvussa esitellään tutkimuksessa käytetty aineisto ja menetelmät. Tutkimuksen tulokset esitellään neljännessä luvussa, ja artikkeli päättyy kokoavaan lukuun.

2 Taustaa

2.1 Diffuusio

Diffuusiolla viitataan kielitieteessä äänteenmuutoksen etenemiseen tai leviämiseen. Sen lopputuloksena voi olla äänteen muuttuminen kokonaan toiseksi, kuten täydellisen assimilaation tapauksessa, jossa viereisen foneemin vaikutus on johtanut äänteenmuutokseen. Esimerkiksi verbimuoto *pääsnyt* on vaihtunut muotoon *päässyt* täydellisen progressiivisen (etenevän) kosketusassimilaation myötä (Hakulinen 2000, 59). Toisaalta diffuusion tuloksena voi olla äänteen katoaminen, kuten tutkimassamme *hd*-tapauksessa. Tällainen prosessi voi edetä hitaasti tai nopeasti, pysähtyä, jatkua uudelleen, koskettaa lopulta koko leksikkoa tai jäädä vain pienemmän sanaryhmän käyttöön. Klassisena esimerkkinä diffuusiosta suomessa on pidetty kielen lipumista hitaasti kohti vokaalivartaloisuutta: esimerkiksi *niemi*-sanan partitiivi on nykykielessä lähes yksinomaan *niemeä* eikä *nientä* (Paunonen 2003, 204).

Kansainvälisessä kielitieteessä on usein puhuttu *leksikaalisesta diffuusiosta* eli äänteenmuutoksen leviämisestä sanasta toiseen (Wang 1969). Suomen kielen agglutinoivaan luonteeseen sopii kuitenkin paremmin ajatus morfologisesta diffuusiosta, joka etenee muotoryhmistä toiseen (Mielikäinen 1995). Aila Mielikäinen on tiivistänyt morfologisen diffuusion merkityksen kolmeen seikkaan: (1) Morfologinen diffuusio ottaa huomioon kielen kaikki tasot eikä vain foneettista tai fonologista, (2) Kuvaustapa sopii sekä vanhoihin aluemurteisiin että nykypuhekieleen, eli eri-ikäiset ja erilaiset kielimuodot tulevat kyseeseen, sekä (3) Morfologista diffuusiota voi soveltaa erityyppisten muutosten kuvaamiseen. Mielikäinen mainitsee diffuusion – sekä morfologisen että leksikaalisen – olevan nimenomaan kuvaustapa, ei varsinainen kielen analyysimenetelmä. Diffuusio ei selitä muutoksen syitä vaan ainoastaan seuraa sen leviämistä. Vaihtelun selityksissä on Mielikäisen mukaan otettava huomioon morfologian lisäksi äänneympäristö, painosuhteet, syntaksi ja sanastoon liittyvät muuttujat. (Mielikäinen 1995, 331–332.)

Kuparisen ja kollegoiden (2022) tutkimuksessa diffuusiota tarkasteltiin leksikaalisen ja morfologisen diffuusion hybridinä eikä vain fennistiikassa tyypilliseen tapaan morfologian näkökulmasta. Keskenään samanlaiset lekseemit yhdistettiin edustamaan samaa muotoryhmää, ja kadon etenemistä tarkasteltiin muotoryhmien välillä ja tarkemmin vielä yksittäisten lekseemien tasolla. Tutkimus osoitti, että katomuodot etenevät yleiskielen *hd*-yhtymässä diffuusiomallin mukaisesti hitaasti muotoryhmistä ja lekseemeistä toisiin, mutta valtaosa katoesiintymistä keskittyy tiettyihin lekseemeihin. Erityisesti katoa näyttävät vetävän puoleensa lukusanat sekä verbit *lähteä* ja *tehdä*. Käsillä oleva artikkeli jatkaa

samoista lähtökohdista, eli diffuusioon vaikuttavista tekijöistä puhuttaessa tarkoitetaan leksikaalisen ja morfologisen diffuusion hybridiä.

2.2 hd-*yhtymä*

Fred Karlsson (2008, 65) esittää, että suomessa on itse asiassa kolme muunnosta *h*:sta: soinniton laryngaalinen [h], soinnillinen laryngaalinen [ɦ] ja soinniton palataalinen [χ]. Ensimmäinen esiintyy soinnittomien konsonanttien edessä, etenkin jos sitä edeltää väljä vokaali (*rahka*, *tähkä*). Toinen esiintyy sanan alussa, vokaalien välissä, ennen soinnillista konsonanttia tai soinnillisen konsonantin jäljessä (*hyvä*, *vaha*, *sähly*, *kulho*). Kolmatta muunnosta esiintyy soinnittomien konsonanttien edessä erityisesti *i*:n jäljessä (*vihko*). Tutkimuksemme *hd*-yhtymän *h* edustaa aina toista muunnosta eli soinnillista laryngaalista [ɦ]:ta, koska sitä seuraa soinnillinen klusiili d (*mahdollinen*, *johdattaa*, *kahdeksikko*) tai *d*:n kadon toteutuessa vokaali (*mahollinen*, *johattaa*, *kaheksikko*). Kummassakin tapauksessa puheessa ääntyy sama [ɦ], kuten Karlsson (mp.) toteaa. Kun *d*:n katoaminen ei aiheuta *h*-foneemiin artikulatorisia muutoksia eikä lekseemiin merkityseroa, mitään selvää estettä muutokselle ei ole. Voi myös olla, että kieli saattaa joskus pyrkiä kohti taloudellista, säästävää ääntämistapaa. Näin on tapahtunut kielihistoriassa aiemminkin: esimerkiksi muutoksessa varhaiskantasuomesta myöhäiskantasuomeen koko joukko vanhoja konsonantteja (9 kpl) katosi lähinnä palautumalla muihin foneemeihin (Lehtinen 2007, 94). Uusia konsonantteja syntyi tuolloin kaksi, joista toinen oli /h/. Hävinneitä foneemiluokkia olivat spirantit, kaikki liudentuneet konsonantit sekä suhusibilantti ja -affrikaatta. (Mts. 96.) Säästävä ääntämistapa on kuitenkin vain yksi katoon mahdollisesti johtavista syistä, eikä yhteen helppoon syyhyn pidä liiaksi tukeutua. Myöhemmässä analyysiosuudessamme selvitämme lisää *d*:n katoon vaikuttavia tekijöitä.

Soinnillinen klusiili *d* lukeutuu suomen kielen foneettisen kentän tuoreimpiin aineksiin. Se on historiallisesti lainaa ruotsista, eikä sitä perinteisesti ole esiintynyt suomen murteissa vaan se on luotu kirjakielen tarpeisiin. Foneemi on ollut osa kieltämme kuitenkin jo vuosisatoja, joskin vielä 1900-luvulla sen lausumista jouduttiin erikseen harjoittelemaan kansakoulussa. Karlsson (2008, 56) nimittää *d*:tä soinnilliseksi apikoalveolaariseksi tai apikodentaaliseksi klusiiliksi. Sen soinniton astevaihtelupari on *t*, joka vanhemmissa lainasanoissa onkin edustanut *d*:tä (esim. *tohtori*, *tanssi*). Uusimmissa lainasanoissa on säilytetty lähtökielen *d* (*demokratia*, *dyyni*). Nykymurteissa *d*:n variantteina on käytössä lähinnä kato ja *r* (Mustanoja 2011, 309).

2.3 Katsaus aiempaan tutkimukseen

Harri Mantila (2004, 327) on todennut, että alun perin itäinen ja pohjoinen katoedustus on selvästi yleistymässä sellaisillekin alueille, joilla on vanhastaan ollut käytössä *t : r* -astevaihtelu. Hän arvelee, että lounaisten, lounaisten välimurteiden, hämäläis- ja eteläpohjalaismurteiden alueella yleisin puhekielinen variantti yleiskielen *d*:n jälkeen olisi jo kato. Mantila mainitsee yleistymisen olleen nopeaa erityisesti *h*:n jäljessä (*mahoton*,

tehä, *lähen*) ja *hd*:llisten numeraalien olevan paikoin jo kokonaan kadollisia (*kahen*, *yheksän*).

Alasatakuntalaismurretta tarkastelleen Kurjen (2005, 107–110) kolmen aikapisteen tutkimuksessa (70-luvun taite, 90-luvun taite ja 2000-luvun taite) alkuperäisen valtavariantti *r*:n osuus *h*:n jälkeisessä asemassa on yhteisössä lähes puolittunut (96 % > 50 %). Suurimman osan *r*:n menettämästä osuudesta on ottanut jo alkujaan seuraavaksi yleisin variantti *d* (4 % > 39 %). Katoa ei 1970-luvun taitteen aineistossa esiintynyt lainkaan, mutta 2000-luvun taitteeseen tultaessa se kattoi 11 prosenttia kaikista tapauksista. Kurjen (mts. 96) mukaan meneillään on siis kaksi voimakasta muutosta: yhtäältä murteellisen tremulantin väistyminen *d*:n tieltä ja toisaalta katovariantin tunkeutuminen kielimuotoon yhä voimakkaammin. Tämä perusasetelma koskee yleiskielen *d*:n vastineita sekä vokaalin jälkeisessä että *h*:n jälkeisessä asemassa, vaikka muutoin äänneympäristö eroja aiheuttaakin.

Liisa Mustanoja (2011) on Tampereen puhekieltä koskevassa kahden aikapisteen reaaliaikatutkimuksessaan (1970-luku ja 1990-luku) selvittänyt muiden piirteiden ohella yleiskielen *d*:n vastineiden edustusta. Hän erotteli tutkimuksessaan *d*:n vastineiden esiintymät neljällä tavalla: *d*, *r*, *ð* ja kato sisältäen mahdolliset siirtymä-äänteet *j*, *v* ja *h*. *hd*-tapauksissa esiintyi kaikkia yleiskielen *d*:n nykyaikaisia vastineita: *lähdin*, *tehrän*, *kahðeksalta*, *kaheksan*. (Mts. 307–309.) Vaikka Mustanoja onkin litteroinut erikseen *r*:n ja puolitremulantti *ð*:n, hän huomauttaa, että kyse on pikemminkin jatkumosta ja eroa on erittäin vaikea kuulla – varsinkin, kun täry on joskus jopa puoli tai puolitoista ja äänteen kestoon vaikuttaa kulloisenkin puhujan idiolektiin tai tilanteiseen vaihteluun kytkeytyvä puhenopeus (Mustanoja–O'Dell 2007). Katomuotojen esiintymisen tulkintaan tämä epätarkkarajaisuus (*ð*:n liukuma *hd*-yhtymässä) ei kuitenkaan vaikuta, koska *ð* esiintyy vaihtoehtoisena niillä puhujilla, jotka käyttävät kadon sijasta lähinnä *r*-varianttia. *d*-variantin käyttö voi olla puhujilla varsin yksinomaista (Mustanoja 2011, 327), ja katovariantin luonne on joko–tai-tyyppinen (mts. 315) ja sikäli helppo tulkita.

Mustanojan havainnot myötäilevät aiempaa tutkimusta siinä, että eniten katoedustusta on *hd*-tapauksissa. Näissä katovarianttia kantavat varsinkin *hd*-yhtymän sisältävät lukusanat sekä *lähteä*-verbi, mutta katovariantin näennäiseen runsauteen aineistossa vaikuttavat usein haastattelupuheessa toistuvat vuosiluvut. Tosiasiassa katovariantin käyttö aineiston puhujilla on kauttaaltaan melko vähäistä. *d*:n ja *r*:n suosijoita aineistossa on suunnilleen yhtä paljon. (Mustanoja 2011, 335.) Mustanojan aineistossa on havaittavissa mielenkiintoinen taipumus, että *r*-variantin käyttö lisääntyy ikääntymisen ja työelämästä jättäytymisen myötä, ja keski-ikäisillä puolestaan esiintyy *d*-varianttia enemmän kuin samoilla henkilöillä nuorempana. Näennäinen *r*-variantin käytön lisääntyminen ei täten merkitse *r*:n ekspansiivisuutta Tampereen kaupunkipuhekielessä vaan kielii lähinnä idiolekteissa tapahtuneesta muutoksesta; idiolektit muuttuvat yhteisön kielen pysyessä suhteellisen stabiilina kahden aikapisteen perusteella. (Mts. 335–336.)

Hanna Lappalainen, Liisa Mustanoja ja Michael O'Dell ovat tutkimuksessaan (2019) erotelleet *hd*:n numeraaleissa, *hd*:n nimissä ja *hd*:n muissa lekseemeissä. Kolmijaon

perusteella kadon esiintymisessä oli merkittävä ero: numeraaleissa puhekielisyyttä eli katoa oli paljon mutta nimissä ei niinkään.

Katsauksessa aiempaan tutkimukseen kävi ilmi, että *hd*-tapauksia on lajiteltu aiemminkin vaikkei juuri käyttämiemme kielensisäisten muuttujien näkökulmasta. Kurki (2005) tutki *d*:tä ja sen vastineita pitkän vokaalin yhteydessä muttei nimenomaan *hd*-yhtymää ja pitkää vokaalia. Astevaihtelun vaikutusta katoon Kurki (2005, 122) on kommentoinut toteamalla, että vartalonsisäisen astevaihtelun ulkopuolella olevilla saneilla on vahvasti katoa suosiva vaikutus ja astevaihtelun alaisilla saneilla puolestaan karttava vaikutus. Hän pitää astevaihtelun merkitystä kadon esiintymisessä vähintään yhtä merkittävänä kuin äänneympäristön, josta hän totesi, että *h*:n jälkeisellä asemalla on katoa paljon suosiva ja vokaalin jälkeisellä asemalla katoa karttava vaikutus. Numeraaleista Kurki on todennut katovariantin taajuuden olevan suurempi tapauksissa *kahdeksan* ja *yhdeksän* kuin *yksi*, *kaksi*, *viisi* tai *kuusi*. (Mp.) Syy piilee jälleen *h*:n jälkeisessä asemassa ja astevaihtelussa.

Kuparinen kollegoineen (2022) selvitti tutkimuksessaan Helsingin puhekielen korpuksesta, onko *hd*-yhtymissä tavattava kato aidosti jähmettynyt lukusanoihin, keskeisiin verbeihin (*tehdä*, *nähdä*, *lähteä*, *ehtiä*, *tahtoa*) ja *mahdollinen*-sanueeseen, kuten Pirkko Nuolijärvi (1986, 112, 114) on esittänyt. Hypoteesina oli, että kato ei olisi jähmettynyt vaan etenisi diffuusiomallin mukaisesti hitaasti leksikon ja morfologisten muotoryhmien sisällä. Lisäksi kirjoittajat olettivat kadon osuuden kasvaneen niissä lekseemeissä, joissa sitä on aiemminkin esiintynyt. Kolmen aikapisteen reaaliaika-aineisto paljasti, että kato on huomattavasti yleistynyt Helsingin puhekielessä 1970-luvulta 2010-luvulle tultaessa: 1990-luvulla syntyneiden puhujien mediaaniarvo kadon osuudelle on yli 80 prosenttia. Kato on siis levinnyt huomattavan pitkälle, ja näennäisaikaennusteen mukaan se saattaisi lopulta korvata *hd*-yhtymän säilyttäneet muodot. Vuosien varrella on tapahtunut kahdenlaista diffuusiota: katomuodossa esiintyviä lekseemejä on tullut lisää niin 1990- kuin 2010-luvuillakin, ja lekseemejä muotoryhmiksi yhdistelemällä kirjoittajat havaitsivat myös morfologista diffuusiota tapahtuneen. Muutoksen syihin ei tutkimuksessa kuitenkaan otettu kantaa, koska kuten kirjoittajat esittävät ”diffuusiomalli ei pyri eikä pysty selittämään muutoksia, vain kuvaamaan niitä” (Kuparinen ym. 2022, 31).

Käsillä oleva artikkeli osaltaan jatkaa katomuotojen etenemisen selvittelyä Helsingin puhekielestä: kun merkittävää muutosta kerran on tapahtunut, mitkä syyt siihen ovat johtaneet? Millaiset muuttujat vaikuttavat kadon todennäköisyyteen? Koska Kuparisen ja kollegoiden (2022) tutkimuksessa tarkasteltiin kadon yleisyyttä vuosikymmenittäin ja ikäryhmittäin, tarkastelemme tässä artikkelissa erityisesti kielensisäisiä muuttujia. Tarkastelemme kadon etenemistä regressiomallien avulla: mitkä yksittäiset muuttujat ja niiden yhdistelmät parhaiten ennustavat katoa.

3 Aineisto ja menetelmät

3.1 Helsingin puhekielen pitkittäiskorpus

Tutkimusaineistona on käytetty Helsingin puhekielen pitkittäiskorpusta (Helpuhe 2014). Korpus on syntynyt alkujaan Helsingin yliopistossa toteutetussa hankkeessa, josta vastasivat pääasiassa Terho Itkonen ja Heikki Paunonen. Kolmatta keruukierrosta on johtanut Hanna Lappalainen. Korpus sisältää haastatteluja kolmelta vuosikymmeneltä: 1970-, 1990- ja 2010-luvulta. Haastatteluihin valittiin informantteja ajan hengessä – labovilainen sosiolingvistiikka oli juuri rantautunut tuoreeltaan Suomeen – kielenulkoisten muuttujien perusteella, joita olivat muun muassa sosiaaliluokka, ikä, asuinpaikka ja sukupuoli. Jatkokierroksilla jatkettiin samalla kaavalla, jotta tutkimustulokset olisivat vertailukelpoisia. Korpus mahdollistaakin paneeli- ja trenditutkimusten tekemisen ja yhdistelemisen, sillä keski-ikäisten ja vanhojen puhujien ryhmä koostuu aina jo aiemmalla kierroksella haastatelluista puhujista.

Haastattelut olivat noin tunnin mittaisia ja käsittelivät arkisia aiheita informanttien elämää koskien sekä asenteita Helsinkiä ja siellä puhuttavia kieliä kohtaan. Helsingin puhekielen tutkimushanke liitettiin osaksi valtakunnallista Nykysuomalaisen puhekielen murros -hanketta vuonna 1976. Mukana tässä hankkeessa olivat myös Tampereen, Turun ja Jyväskylän kaupunkipuhekielet. Sittemmin kolme keruukierrosta on toteutettu Helsingin lisäksi vain Tampereella, mutta Helsingin ja Tampereen aineistot eivät ole täysin vertailukelpoisia, sillä Tampereella ei haastateltu 1990-luvulla uutta nuorten ryhmää. Tampereen korpuksen kolmas kierros ei ole myöskään vielä täydellisen valmis. (Paunonen 2009; Mustanoja 2011.)

Korpus on käytettävissä CSC:n Kielipankissa ja relevantti; onhan siihen koottu pääkaupunkimme puhekieltä puolen vuosisadan takaa näihin päiviin asti. *hd*-yhtymän diffuusion syitä siitä ei ole tätä ennen tällä tarkkuudella tarkasteltu. Tässä tutkimuksessa käytetään Helsingin puhekielen pitkittäiskorpuksesta 199 haastattelua, joista oli saatavilla litteraatti tutkimusta aloitettaessa. Tutkimukseen on poimittu aineiston litteraateista kaikki *hd*-tapaukset (N = 5537), esiintymien laatu (*d*, N = 3315; kato, N = 2222) ja esiintymislekseemit (yhteensä 184). *hr*-tapaukset on jätetty huomiotta, koska niiden määrä on verrattain vähäinen (N = 14). Aineiston rakenne vuosikymmenittäin on esitetty taulukossa 1. Lisäksi taulukossa on esitetty *hd*-yhtymäesiintymien laatu ja määrä. Määristä huomataan selvästi, että nuoret käyttävät katomuotoja runsaammin kuin vanhemmat ryhmät ja että katomuodot yleistyvät ajan kuluessa. Yhtymän ikä- ja vuosikymmenkohtaista variaatiota tarkastellaan lähemmin toisaalla (Kuparinen ym. 2022).

Taulukko 1. Aineiston jakautuminen kohortteihin ja haastattelukierroksiin sekä *hd*-yhtymäesiintymien laatu ja määrä.

Kohortti	Haastattelut 1972–1974 *d* = 1966, kato = 841	Haastattelut 1991–1992 *d* = 738, kato = 681	Haastattelut 2013 *d* = 611, kato = 699
Kohortti 1 s. 1900–1907	65–74 vuotta 46 puhujaa *d* = 731, kato = 161		
Kohortti 2 s. 1927–1932	40–47 vuotta 41 puhujaa *d* = 776, kato = 190	59–65 vuotta 9 puhujaa *d* = 349, kato = 82	81–86 vuotta 4 puhujaa *d* = 117, kato = 30
Kohortti 3 s. 1952–1957	15–22 vuotta 39 puhujaa *d* = 459, kato = 490	34–40 vuotta 10 puhujaa *d* = 263, kato = 192	56–61 vuotta 9 puhujaa *d* = 210, kato = 167
Kohortti 4 s. 1971–1975		16–21 vuotta 14 puhujaa *d* = 126, kato = 407	38–42 vuotta 14 puhujaa *d* = 194, kato = 247
Kohortti 5 s. 1994–1997			16–19 vuotta 13 puhujaa *d* = 90, kato = 255

3.2 Logistinen regressio ja päätöspuu

Aineiston analyysi toteutettiin kahdella eri tilastotieteellisellä menetelmällä, logistisella regressiolla (ks. esim. Kaakinen–Ellonen) ja päätöspuulla (Breiman–Friedman–Olshen–Stone 1984). Molemmat menetelmät pyrkivät selvittämään, miten selittävät tekijät vaikuttavat selitettävän muuttujan (*hd*-esiintymä) todennäköisyyteen. Logistinen regressio on tavallisen regressiomallin erikoistyyppi, jossa selitettävä muuttuja on kategorinen kahdella arvolla. Malli ennustaa todennäköisyyttä sille, että selitettävän muuttujan arvo on toinen kahdesta kategoriasta, tässä siis toteutuuko kato vai ei. Logistinen regressio antaa yleiskuvan siitä, miten selittävät muuttujat vaikuttavat selitettävän muuttujan todennäköisyyteen, mutta se ei juuri kerro muuttujien keskinäisestä suhteesta. Siksi olemme käyttäneet myös päätöspuuta, jonka etuna on, että se tuottaa havainnollisen puukuvion, jonka avulla selittävien muuttujien arvoissa tapahtuvien erojen vaikutusta selitettävään muuttujaan on helppo seurata.

Käyttämässämme päätöspuuanalyysissa (Hothorn–Hornik–Zeileis 2006) aineistoa jaetaan vaiheittain osiin siten, että kukin jako tehdään sillä selittävällä muuttujalla, jolla on selkein vaikutus selitettävään muuttujaan. Kullekin mahdolliselle selittävälle muuttujalle testataan nollahypoteesia, jonka mukaan muuttujan arvolla ei ole vaikutusta selitettävän muuttujan esiintymiseen. Jako tehdään sillä muuttujalla, jolla nollahypoteesi hylätään selkeimmin. Jakoa jatketaan, kunnes nollahypoteesia ei voida hylätä eli tilastollisesti merkitsevää eroa eri osien välillä ei enää saavuteta. Tästä vaiheittaisesta jaosta voidaan piirtää

puukuvio, jota luetaan ylhäältä alaspäin. Kaikkein merkittävin jako tapahtuu ylhäällä, puun ensimmäisessä haarassa, ja alimmaisina ovat puolestaan niin sanotut lehtisolmut, joissa annetaan selitettävälle muuttujalle todennäköisyys. Mallit täydentävät täten toisiaan hyvin: logistinen regressio antaa yleiskuvan muuttujien vaikutuksesta siinä missä päätöspuu kuvaa muuttujien välistä suhdetta.

Suomalaisessa sosiolingvistiikan tutkimuksessa päätöspuuta on aiemmin hyödyntänyt muun muassa Katri Priiki väitöskirjassaan (2017; ks. myös esim. Väänänen 2016). Priiki (2017, 39) mainitsee, että päätöspuut ovat yleisessä käytössä koneoppimisessa ja lääketieteessä, ja kielitieteessä menetelmä on ollut käytössä viime vuosikymmenten varrella jonkin verran. Myös Priiki pitää päätöspuun merkittävänä etuna sitä, että se tuo näkyviin selittävien muuttujien välisiä yhteyksiä ja sitä, että puukuvio on logistisen regression tuottamiin kerroinlukuihin verrattuna jokseenkin helppolukuinen. Toisaalta Priiki mainitsee puuanalyysin heikkoudeksi sen, ettei ensimmäisen haarautumisen jälkeen tulevista muuttujista enää saada yleiskuvaa. Siksi hän on omassa tutkimuksessaan käyttänyt myös satunnaismetsämenetelmää, joka tuottaa haarautuvien puukuvioiden sijasta tunnuslukuja, jotka ilmaisevat kunkin muuttujan tärkeyttä tarkasteltavan vaihtelun ennustamisessa (Priiki 2017, 39–40; ks. myös Ho 1995). Koska meidän tutkimuksessamme on vain yksi selitettävä muuttuja (*hd*-esiintymä), saamme tarpeeksi tietoa käyttämällä logistista regressiota ja päätöspuuta yhdessä.

3.3 Tilastomenetelmissä käytetyt muuttujat

Aineistossa on *hd*-esiintymiä kaikkiaan 5537, joista suunnilleen 40 % esiintyy kadollisena (*mahoton*) ja 60 % *d*:llisenä (*mahdoton*). Aineiston perusteella lähdimme selvittämään, mitkä kielelliset seikat vaikuttavat kadon todennäköisyyteen *hd*-yhtymissä. Tarkastelussa ei ole otettu huomioon puhujiin liittyviä muuttujia, sillä niitä on käsitelty samasta aineistosta toisaalla (Kuparinen ym. 2022; ks. myös Lappalainen ym. 2019). Tässä luvussa esitellään kaikki kuusi muuttujaa, joiden pohjalle analyysi luvussa 4 on rakennettu.

1. **Frekvenssi** on jatkuva-arvoinen muuttuja, joka kertoo, kuinka monta kertaa *hd*-yhtymän sisältävä lekseemi esiintyy aineistossa. Frekvenssi on laskettu vain sellaisista tämän aineiston tapauksista, joissa *hd*-yhtymä voisi esiintyä. Niinpä mukana on esimerkiksi passiivi *lähdetään*, mutta A-infinitiivi *lähteä* on jätetty pois.

2. **Painoasemalla** tarkoitetaan sitä, esiintyykö *hd* painollisen tavun (*kahdeksan*) vai painottoman tavun (*pyörähdä-*[1]) jäljessä.

3. **Vokaali** saa arvon lyhyt, pitkä tai diftongi *hd*-yhtymää edeltävän vokaalin laadun mukaan.

4. **Astevaihtelu** saa arvon ”kyllä” tai ”ei” sen mukaan, kuuluuko sana edelleen astevaihteluun (*kahta* : *kahden*) vai ei (*kahdeksan*).

1 Tutkimuksessa käytetään sellaisia muotoja, joissa *d* esiintyy.

5. Selittävä muuttuja **vartalo** voi saada kolme arvoa sen mukaan, missä vartalossa *hd*-yhtymä esiintyy. Vaihtoehdot ovat vokaalivartalo (*lähdetään*), konsonanttivartalo (*tehdään*) tai molemmat (*mahdollista*, *mahdollisen*)[2]. Tällainen tarkastelu antaa tarkemman kuvan astevaihtelun vaikutuksesta katomuotojen esiintymiseen.

6. Viimeinen diffuusiota selittävä muuttuja on **merkitys**, joka on jaettu kolmeen joukkoon: nimiin, *yhdessä*-pesueeseen (sanoihin, joiden kantana on sana *yksi* mutta jotka ovat merkitykseltään etääntyneet kantasanasta) ja muihin.

Taulukossa 2 on esitetty *d*-esiintymien ja katomuotojen absoluuttiset sekä suhteelliset osuudet kullekin muuttujan arvolle. Frekvenssi on selkeyden vuoksi jaettu kahteen joukkoon käyttäen rajana 200 esiintymää. Taulukon arvoja hyödynnetään luvussa 4.1 logistisen regression vertailuarvojen valinnassa: vertailuarvona käytetään sitä muuttujan arvoa, joka esiintyy aineistossa useimmin.

Taulukko 2. Absoluuttiset ja suhteelliset arvot *d*- ja katoesiintymille kullakin muuttujan arvolla.

Muuttuja	Arvo	*d*-esiintymät	Katoesiintymät	Yhteensä
Frekvenssi	> 200	1425 (43 %)	1901 (57 %)	3326
	≤ 200	1890 (85 %)	321 (15 %)	2211
Painoasema	Painoton	101 (97 %)	3 (3 %)	104
	Painollinen	3214 (59 %)	2219 (41 %)	5433
Vokaali	Pitkä	151 (99 %)	1 (1 %)	152
	Diftongi	137 (89 %)	17 (11 %)	154
	Lyhyt	3027 (58 %)	2204 (42 %)	5231
Astevaihtelu	Kyllä	2158 (62 %)	1313 (38 %)	3471
	Ei	1157 (56 %)	909 (44 %)	2066
Vartalo	Konsonanttivartalo	648 (69 %)	297 (31 %)	945
	Vokaalivartalo	2150 (55 %)	1772 (45 %)	3922
	Molemmat	517 (77 %)	153 (23 %)	670
Merkitys	Nimet	301 (96 %)	13 (4 %)	314
	Yhdessä	281 (88 %)	39 (12 %)	320
	Muut	2733 (56 %)	2170 (44 %)	4903

2 Suomen kaikilla sanoilla on vokaalivartalo, mutta osalla sanoista on lisäksi konsonanttivartalo. Tällaisia sanoja kutsutaan kaksivartaloisiksi. Konsonanttivartalo on aina heikkoasteinen, mikäli vartalossa esiintyy astevaihtelua. Tämä seikka on otettava huomioon, sillä tutkimuskohteemme *hd* on *ht*-yhdistelmän heikko astevaihtelupari. Konsonanttivartalot ovat ennen olleet yleisempiä, mikä on havaittavissa esimerkiksi vanhasta suomalaisesta kirjallisuudesta. Konsonanttivartaloiden vähenemiseen vuosikymmenten saatossa on vaikuttanut pyrkimys yhteneväisempään paradigmaan.

4 Aineiston analyysi ja tulokset

Aineiston analyysi toteutetaan luvussa 3.2 esitellyillä logistisella regressiolla ja päätöspuulla. Menetelmät on valittu siten, että ne täydentävät toisiaan. Logistinen regressio antaa kokonaiskuvan muuttujien vaikutuksesta kadon todennäköisyyteen: kukin muuttuja saa kertoimen sen mukaan, mikä muuttujan vaikutus on kadon todennäköisyyteen verrattuna perusarvoon. Logistinen regressio siis paljastaa, onko jonkin muuttujan vaikutus kadon todennäköisyyteen positiivinen vai negatiivinen. Päätöspuu puolestaan kertoo muuttujien yhteisvaikutuksesta: kuinka paljon katoa esiintyy esimerkiksi tietyn frekvenssin konsonanttivartaloisissa sanoissa? Koska logistinen regressio tarjoaa yleiskuvan kaikista muuttujista, aloitamme analyysimme siitä luvussa 4.1. Tätä kokonaiskuvaa täydennetään luvussa 4.2 päätöspuuanalyysilla.

4.1 hd-*yhtymän analyysi logistisella regressiolla*

Logistinen regressio ennustaa selitettävän muuttujan esiintymää selittävien muuttujien avulla. Tässä tapauksessa selitettävä muuttuja on *hd*-esiintymä (kato tai *d*), ja selittävät muuttujat esiteltiin luvussa 3.3. Logistisessa regressiossa valitaan kullekin kategoriselle muuttujalle vertailuarvo, johon muuttujien muita arvoja verrataan. Käytämme selitettävän muuttujan (*hd*-esiintymä) vertailuarvona tapauksia, joissa *d* esiintyy (esim. *kahdeksan*). Näin ollen logistinen regressio pyrkii ennustamaan kadon esiintymistä (*kaheksan*) selittävien muuttujien avulla. Selittävien muuttujien osalta vertailuarvot valittiin puolestaan niiden yleisyyden mukaan (ks. taulukko 2). Vertailuarvot ovat seuraavat: painoasema painollisen tavun jäljessä, edeltävä vokaali on lyhyt (ei pitkä tai diftongi), *hd* esiintyy vokaalivartalossa, sana on astevaihtelussa eikä kuulu erityisiin merkitysryhmiin (tällainen sana on esimerkiksi *kahde-*). Näin ollen logistinen regressio ennustaa tapauksia, joissa kato esiintyy ja selittävä muuttuja poikkeaa yllä olevista vaihtoehdoista.

Kukin selittävä muuttuja saa logistisessa regressiossa kertoimen sen mukaan, mikä sen arvon vaikutus on selitettävään muuttujaan. Näin ollen positiivinen kerroin lisää kadon todennäköisyyttä verrattuna vertailuarvoon. Kertoimien lisäksi taulukossa 1 esitetään myös kertoimien luottamusväli (2,5–97,5 %) sekä p-arvo kullekin muuttujalle. Kertoimien p-arvot on laskettu Waldin testillä (Wald 1943), jonka nollahypoteesina on, että selittävän muuttujan arvolla ei ole vaikutusta selitettävän muuttujan arvoon, ja luottamusvälit ovat Wald-luottamusvälejä. Merkitsevyyden taso on $p < 0{,}05$, ja merkitsevyydet on laskettu Holm–Bonferroni-korjausta (Holm 1979) käyttäen. Korjauksen tavoitteena on välttää vääriä positiivisia tuloksia, jotka johtuvat siitä, että testejä tehdään useita. Siinä kaikkien testien p-arvot järjestetään pienimmästä suurimpaan. Testejä tarkastellaan tässä järjestyksessä niin, että haluttu merkitsevyystaso jaetaan testien lukumäärällä: kun tässä tapauksessa tavoitellaan merkitsevyystasoa 0,05 ja testejä on 10, pienintä p-arvoa voi pitää merkitsevänä vasta, jos se on alle $0{,}05/10 = 0{,}005$. Seuraavan testin kohdalla tarkastellaan tilannetta, jossa testien lukumäärä on yhtä pienempi, joten toiseksi pienimmän p-arvon on oltava alle $0{,}05/9 = 0{,}00556$ ja niin edelleen, kunnes

viimeisen testin p-arvolle rajana on 0,05/1. Kaikki malliin valitut selittävät muuttujat ovat tilastollisesti selvästi merkitseviä.

Taulukko 3. hd-yhtymän esiintymien selittävien tekijöiden kertoimet, luottamusvälit ja p-arvot. Esiintymän vertailuarvona on käytetty arvoa *hd*. Vertailuarvoina muuttujille ovat painoasema painollisen tavun jäljessä, lyhyt edeltävä vokaali, *hd* esiintyy vokaalivartalossa, sana on astevaihtelussa, sana ei kuulu erityisiin merkitysryhmiin. N = 5537.

	Kerroin	2,5 %	97,5 %	p-arvo
Vakiotermi	–1,355	–1,532	–1,178	< 0,001
Frekvenssi (jatkuva)	0,003	0,003	0,004	< 0,001
Painoasema: painoton	–2,141	–3,300	–0,982	< 0,001
Vokaali: pitkä	–3,737	–5,699	–1,752	< 0,001
Vokaali: diftongi	–0,788	–1,317	–0,259	< 0,01
Vartalo: konsonanttivartalo	–1,065	–1,244	–0,887	< 0,001
Vartalo: molemmat	–0,520	–0,763	–0,279	< 0,001
Astevaihtelu: ei	0,333	0,177	0,489	< 0,001
Merkitys: nimet	–1,870	–2,450	–1,290	< 0,001
Merkitys: yhdessä	–1,304	–1,670	–0,938	< 0,001

Selittävistä muuttujista vain frekvenssi ja astevaihteluun kuulumattomuus saavat positiivisen arvon. Näin ollen frekvenssin kasvaessa myös katomuotojen todennäköisyys kasvaa, ja katomuodot ovat todennäköisempiä sanoissa, jotka eivät kuulu astevaihteluun. Näitä kahta muuttujaa lukuun ottamatta kaikki muut muuttujat saavat logistisessa regressiossa negatiiviset kertoimet, mikä tarkoittaa, että kaikki muuttujat taulukossa kuvatulla arvolla vaikuttavat katomuotojen esiintymisen todennäköisyyteen negatiivisesti verrattuna vertailuarvoon. Negatiivisten kertoimien suuruus kertoo, kuinka voimakkaasti arvot vastustavat katoa. Näin siis esimerkiksi *hd*-yhtymän esiintyminen diftongin jäljessä on kadolle epäotollisempi paikka kuin lyhyen vokaalin jäljessä (kerroin –0,788), mutta kuitenkin parempi kuin pitkän vokaalin jäljessä (–3,737). Erityisesti esiintyminen painottoman tavun jäljessä (*unohda-*), pitkän vokaalin jäljessä (*viihdy-*) tai proprinen merkitys näyttävät merkittävästi vähentävän kadon esiintymistä.

On muistettava, että logistisessa regressiossa kaikilla muuttujilla on vertailuarvo. Näin ollen tapaukset, joissa *hd*-yhtymä esiintyy painollisen tavun ja lyhyen vokaalin jäljessä ja vain vokaalivartalossa toimivat mallin perusarvoina, joihin taulukossa 1 esiintyviä arvoja verrataan. Taulukon 1 perusteella voidaan siis päätellä vain näistä perusarvoista poikkeavien arvojen vaikutusta. Taulukon perusteella voidaan todeta, että ainoastaan lekseemin frekvenssin kasvu ja kuulumattomuus astevaihteluun vaikuttaa kadon esiintymiseen

positiivisesti, kun taas kaikki muut morfologiset ja semanttiset eroavaisuudet vaikuttavat negatiivisesti. Logistinen regressio kuitenkin paljastaa myös muuttujien keskinäisiä eroja: osa muuttujien arvoista vaikuttaa selvästi negatiivisemmin kadon todennäköisyyteen (esim. edeltävä pitkä vokaali) kuin toiset (esim. edeltävä diftongi). Muuttujien yhteisvaikutusta tarkastellaan päätöspuuanalyysissä.

4.2 hd-*yhtymän analyysi päätöspuulla*

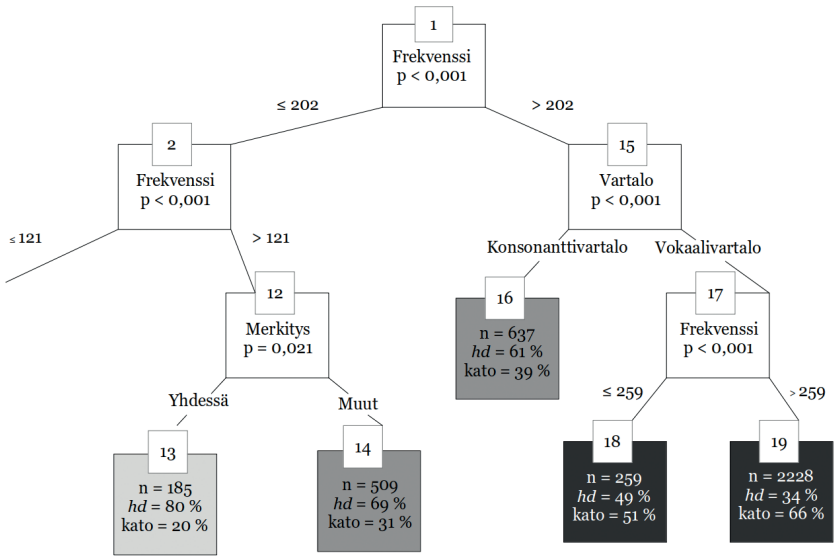
Käyttämässämme päätöspuumallissa (Hothorn ym. 2006) aineisto jaetaan vaiheittain osiin selittävien muuttujien avulla. Jako tapahtuu kahdessa osassa:

> 1. Malli testaa kunkin selittävän muuttujan osalta nollahypoteesia, jonka mukaan selittävän muuttujan arvolla ei ole vaikutusta selitettävän muuttujan esiintymiseen. Tässä tutkimuksessa tarkastellaan siis esimerkiksi sitä, onko *hd*-yhtymän painoasemalla vaikutusta kadon todennäköisyyteen. Mikäli nollahypoteesia ei hylätä, jakoa ei tehdä, ja päätöspuu on valmis.
>
> 2. Mikäli taas nollahypoteesi voidaan yhden tai useamman selittävän muuttujan arvolla hylätä, jako tehdään sen muuttujan perusteella, jonka avulla tehty jako tuottaa selkeimmän eron aineiston eri osien välille eli hylkää nollahypoteesin pienimmällä p-arvolla. Aineiston jakamista jatketaan samalla periaatteella, kunnes nollahypoteesia ei enää hylätä ja puu on valmis.

Edellä kuvatun vaiheittaisen jaon perusteella voidaan piirtää itse päätöspuu ylhäältä alaspäin niin, että kussakin jakovaiheessa puu haarautuu. Haarat johtavat joko uusiin haaroihin tai lehtisolmuihin, joissa esitetään *hd-* ja katomuotojen osuus kyseisessä solmussa. Tässä päätöspuussa nollahypoteesi hyväksytään tai hylätään perustuen permutaatiotestiin (Strasser–Weber 1999). Tilastollisen merkitsevyyden tasona käytetään päätöspuussakin arvoa $p < 0{,}05$, jonka lisäksi hyödynnetään Šidák-korjausta. Yllä kuvatun päätöspuun olemme tehneet R-sovelluksen (R Core Team 2021) lisäosalla *party* (Hothorn–Hornik–Zeileis 2021). Päätöspuun konkordanssiarvo (C) on 0,78, joka kertoo puun ennustavan katoa kohtuullisen hyvin (Baayen–Endresen–Janda–Makarova–Nesset 2013; Priiki 2016, 118).

Olemme luettavuuden vuoksi jakaneet päätöspuun kahteen osaan (kuvio 1 ja kuvio 2). Päätöspuu tulee käsittää ikään kuin puusta riippuvaksi oksaksi, josta erkanee uusia oksia. Ylimmäisenä on noodi, jonka sisällä on koko aineisto. Sitten noodit joko haarautuvat tai päättyvät. Haarautuvat noodit on esitetty valkoisella pohjalla, ja niiden sisällä on sen muuttujan nimi, jonka perusteella jako on tehty. Sen lisäksi noodissa on permutaatiotestin tuottama p-arvo tehdylle jaolle. Noodista lähtevien haarojen oheen on kirjoitettu jaon muodostavat muuttujan arvot. Esimerkiksi tämän päätöspuun noodissa 2 tehdään jako sanan frekvenssin perusteella. Noodista lähtee kaksi haaraa: sanat, joiden frekvenssi on yli 121 ja sanat, joiden frekvenssi on korkeintaan 121. Tämän noodin perusteella myös kuvioiden välinen jako on tehty: kuviossa 2 on päätöspuun se osa, jossa esiintymien frekvenssi on korkeintaan 121.

Päättyvistä noodeista eli lehtisolmuista ei enää synny uusia haaroja. Tällaiset noodit on päätöspuussa esitetty harmaan sävyin. Jokaisen päättyvän noodin ohessa on esitetty tiedot esiintymien kokonaismäärästä ja *hd*-muotojen sekä katomuotojen osuudesta kyseisessä noodissa. Jos katomuotojen määrä ylittää 50 % kaikista esiintymistä, noodin pohjaväri on musta. Jos katomuotoja on noodissa 30–50 %, tausta on tummanharmaa. Muissa tapauksissa taustaväri on vaaleanharmaa.

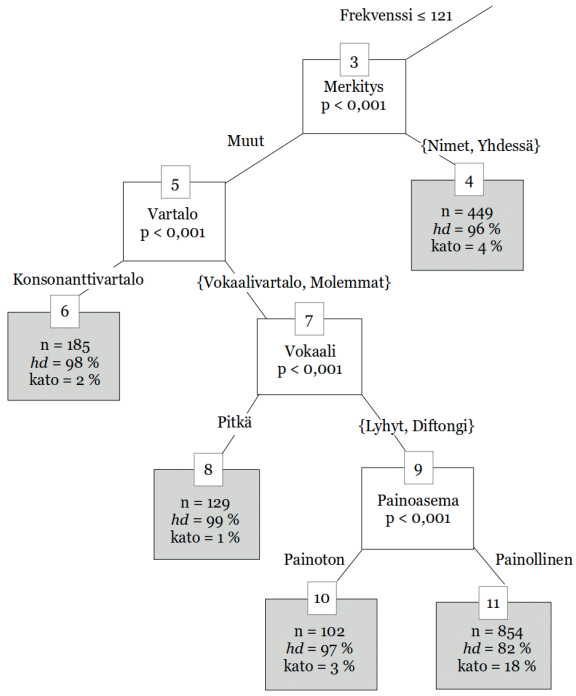


Kuvio 1. Päätöspuu *hd*-yhtymän esiintymien jakaantumisesta. Lekseemit, joiden frekvenssi on aineistossa yli 121. Haarassa 12 tehdään jako vain niiden merkitys-muuttujan arvoille, jotka esiintyvät aineiston siinä haarassa, samoin haarassa 15 vain vartalon niille arvoille, jotka esiintyvät siinä haarassa.

Kuviossa 1 on esitetty ensimmäinen osa päätöspuusta. Ensimmäinen jako on tehty sen perusteella, onko lekseemin frekvenssi yli 202 vai korkeintaan 202. Näin ollen frekvenssi on myös päätöspuuanalyysissä tärkein muuttuja. Tarkastelemme ensin kaikkein tiheimmin esiintyviä lekseemejä. Noodi 15 haarautuu sen mukaan, millainen on lekseemin vartalomorfeemi: esiintyykö *hd*-yhtymä siis vokaalivartalossa vai konsonanttivartalossa. Vokaalivartaloiset tapaukset haarautuvat vielä uudelleen frekvenssin mukaan, mutta tarkastelemalla noodeja 16, 18 ja 19 voidaan havaita jo vartalomorfeemin selvä vaikutus. Korkean frekvenssin sanoissa konsonanttivartaloiset lekseemit esiintyvät harvemmin katomuodoissa (39 %) kuin vokaalivartaloiset lekseemit (51 % ja 66 % frekvenssistä riippuen). Konsonanttivartaloisuus siis vaikuttaa hidastavan katomuotojen etenemistä. On kuitenkin huomattava, että yli 202 esiintymän frekvenssiä ei konsonanttivartaloisista sanoista saa kuin verbi *tehdä*, joka siis kattaa kaikki noodin 16 esiintymät. Toisaalta verbin todistusarvo on frekvenssin osalta melko vahva, koska se on koko aineiston yleisin lekseemi (N = 637).

Kuvion 1 toinen haara esittää tapaukset, jotka esiintyvät aineistossa korkeintaan 202 kertaa. Noodi 2 haarautuu jälleen frekvenssin mukaan tapauksiin, jotka esiintyvät korkeintaan 121 kertaa (kuviossa 2) ja sen yli esiintyviin. Lekseemeissä, jotka esiintyvät yli 121 kertaa, tärkeimmäksi muuttujaksi nousee merkitys. Mikäli lekseemi kuuluu *yhdessä*-sanapesueeseen, on katomuotojen osuus pienempi (20 %) kuin muissa lekseemeissä (31 %). Yhteenvetona kuviosta 1 voidaan siis todeta, että frekvenssin vaikutus on muuttujista kaikkein suurin, mutta sen vaikutusta hidastavat eniten konsonanttivartaloisuus ja merkitystekijät.

Kuviossa 2 on esitetty päätöspuun se puoli, joka haarautui noodista 2 frekvenssin perusteella, eli sellaiset lekseemit, jotka esiintyvät aineistossa korkeintaan 121 kertaa. Ensimmäisenä on kiinnitettävä huomiota siihen, että kaikki päättyvät noodit ovat väriltään vaaleanharmaita eli katomuotojen osuus ei missään noodissa ole yli 30 prosenttia. Näyttää siis selvältä, että matala frekvenssi estää katomuotojen etenemistä. Frekvenssin jälkeen tärkein muuttuja on merkitys, joka haarautuu noodissa 3 tapauksiin, joissa lekseemi kuuluu joko nimien tai *yhdessä*-sanapesueen joukkoon (katomuotoja 4 %) sekä muihin tapauksiin.



Kuvio 2. Päätöspuu *hd*-yhtymän esiintymien jakaantumisesta. Lekseemit, joiden frekvenssi on aineistossa korkeintaan 121.

Merkityksen jälkeen puu haarautuu jälleen vartalon mukaan: konsonanttivartaloiset tapaukset eroavat niistä tapauksista, joissa *hd*-yhtymä esiintyy vokaalivartalossa tai molemmissa vartaloissa. Konsonanttivartaloisten tapauksessa katomuodot ovat todella harvinaisia (2 %). Ei-konsonanttivartaloiset esiintymät haarautuvat vielä edeltävän vokaalin laadun (noodi 7) ja painoaseman (noodi 9) mukaan. Näiden osalta harvinaisemmissa tapauksissa (*hd* pitkän vokaalin tai painottoman tavun jäljessä) katotapauksia ei juuri esiinny (1 % ja 3 %).

Korkeintaan 121 kertaa aineistossa esiintyvistä lekseemeistä kaikkein korkein katomuotojen osuus on noodissa 11. Tähän noodiin kuuluvat sellaiset lekseemit, jotka kuuluvat merkitykseltään joukkoon ”muut”, joissa *hd*-yhtymä ei esiinny konsonanttivartalossa, joissa edeltävä vokaali on lyhyt ja joissa edeltävä tavu on painollinen. Näissä lekseemeissä, joihin kuuluvat esimerkiksi sellaiset kuin *ehdi-* ja *lehde-*, katomuotojen osuus on 18 %. Vaikuttaa siis siltä, että tällainen lekseemin rakenne on katomuotojen osalta prototyyppinen: ne seikat, jotka hidastivat korkean frekvenssin sanojen esiintymistä katomuodoissa, toimivat samoin myös matalan frekvenssin sanoissa.

Päätöspuusta on huomattava myös se, mitä siinä ei ole: malli ei ole hyödyntänyt haarautumisissa muuttujaa astevaihtelu, vaikka logistisen regression mukaan tämäkin muuttuja oli merkitsevä. Astevaihtelu kytkeytyy melko vahvasti vartalovokaalin vaihteluun. Niissä tapauksissa, joissa *hd*-yhtymä esiintyy sekä vokaali- että konsonanttivartalossa, se ei osallistu astevaihteluun (esim. *mahdollinen*). Sen sijaan kaikki konsonanttivartaloiset osallistuvat astevaihteluun kuten myös valtaosa vokaalivartaloisista. Vaikuttaa siis siltä, että astevaihtelun sijaan ensisijaista katomuotojen esiintymisessä onkin, missä vartalossa *hd*-yhtymä esiintyy.

5 *Lopuksi*

Olemme tarkastelleet tässä tutkimuksessa yhteensä kuutta *hd*-yhtymän katomuotojen diffuusiota selittävää tekijää: frekvenssiä, painoasemaa, edeltävää vokaalia, astevaihtelua, vartaloa ja merkitystä. Havaintomme kadon esiintymisen syistä perustuvat tilastotieteellisillä menetelmillä (logistisella regressiolla ja päätöspuulla) tehtyihin löydöksiin aineistosta.

Tutkimuksessa havaittiin, että frekvenssin vaikutus kadon diffuusiossa on keskeisin. Kun lekseemi esiintyy tiheästi, myös kato leviää nopeasti ja laajalle. Taajuusherkkyyden hypoteesin mukaan fonologisehtoiset uudennokset tarttuvat ensimmäisenä frekventeimpiin lekseemeihin (Nahkola 1987, 43; Suihkonen 1992, 45). Kato on suurimmillaan silloin, jos sana on yleinen. Mielenkiintoista on myös sanan vartalon vaikutus kadon diffuusioon. Tutkimuksessa todettiin, että jos *hd* sijaitsee konsonanttivartalossa, se vähintään hidastaa kadon etenemistä.

Selvästi yleisintä kato on vokaalivartaloisissa korkean frekvenssin lekseemeissä (*lähde-*). Seuraavaksi yleisintä kato on, jos sana on konsonanttivartaloinen ja yleinen (*tehdä*-verbi). Vaikuttaa siltä, että matala frekvenssi estää tehokkaasti katomuotojen

etenemistä. Harvinaisinta kato on konsonanttivartaloisilla matalan frekvenssin lekseemeillä. Vartalomorfeemin vaikutus on selvä: korkean frekvenssin sanoissa vokaalivartaloiset lekseemit esiintyvät useammin katomuodoissa kuin konsonanttivartaloiset. *hd*-yhtymän esiintyminen konsonanttivartalossa siis vähentää kadon todennäköisyyttä, vaikka sanalla olisi korkea frekvenssi. Koko aineiston yleisin lekseemi (N = 637) on konsonanttivartaloinen *tehdä*, joka kattaa kaikki päätöspuun noodin 16 esiintymät. Pelkkään frekvenssiin perustuen *tehdä*-verbin sanueiden pitäisi siis esiintyä valtaosin katomuodossa, mutta konsonanttivartaloisuuden tähden kadon osuus on vain 39 % ja *hd*-yhtymän säilyttäneitä muotoja on 61 %.

Vartalovaihtelussa tarkastelimme vain sitä, missä vartaloissa *hd*-yhtymä esiintyy. Näin ollen esimerkiksi numeraalit *yhde-*, *kahde-*, *kahdeksa-*, *yhdeksä-* esiintyvät tutkimuksessa vain vokaalivartaloisina, vaikka *yhde-* ja *kahde-* ovat tosiasiassa kaksivartaloisia. Tulosten perusteella tällä valinnalla ei näytä olevan suurta vaikutusta: molemmat numeraalityypit ovat hyvin tavallisia katomuodoissa. Kuparisen ja kollegoiden (2022) mukaan *kahdeksa-* ja *yhdeksä-* ovat kuitenkin yleisempiä katomuodoissa kuin niiden vastinparit *kahde-* ja *yhde-* (vrt. myös Kurki 2005, 122). On siis myös mahdollista, että sanan kuuluminen ylipäänsä vartalovaihtelun piiriin vähentää kadon todennäköisyyttä, mutta tämän tutkimuksen asetelman perusteella tätä ei voida vahvistaa.

Keskeiseksi katoa estäväksi muuttujaksi aineistossa muodostui myös merkitys. Erisnimet ja *yhdessä*-sanapesue (*yhdistelmä*, *yhdyskunta* jne.) ikään kuin välttelevät katoa, kun taas muissa lekseemeissä kato pääsee leviämään vapaammin. Vahvistimme myös aiemman tutkimuksen (Kuparinen ym. 2022) oletuksen siitä, että kadon diffuusiota vaikuttaa hidastavan esiintyminen painottoman tavun tai pitkän vokaalin jäljessä (*pyörähdä-*, *hiihdä-*).

Yhdistelemällä logistisen regression ja päätöspuun tuloksia voidaan päätellä, että kaikkein prototyyppisin lekseemi katomuotojen esiintymiselle on sellainen, jossa *hd*-yhtymä esiintyy lyhyen vokaalin ja painollisen tavun jäljessä vokaalivartalossa (vain tai konsonanttivartalon lisäksi) eikä katomuodolle ole merkityksen tuomaa estettä. Nämä ehdot pätevät päätöspuuanalyysin mukaan niin korkean kuin matalan frekvenssin sanoihin. Frekvenssi säätelee kadon laajuutta: korkean frekvenssin sanoissa voi esiintyä runsaasti katoa, vaikka edelliset ehdot eivät täyttyisi (esim. *tehdä*-verbi), kun taas matalan frekvenssin sanoissa ehtojen täyttyminen voi antaa mahdollisuuden kadolle, joka ei kuitenkaan ole yhtä laajaa kuin korkean frekvenssin sanoissa (esim. *ehdi*-verbi). Ehdot täyttäviä korkean frekvenssin sanoja ovat esimerkiksi lukusanat ja *lähde*-verbi, joiden onkin jo pitkään havaittu esiintyvän runsaasti katomuotoisina.

Katomuotojen tarkastelu paljastaa erilaisia kielenmuutoksen leviämiseen vaikuttavia tekijöitä. Diffuusio on toisaalta leksikaalista, toisaalta morfologista ja tiettyihin muotoryhmiin keskittyvää. Näiden ehtojen täyttyminen sinänsä riittää muutoksen etenemiseen, mutta sen nopeuteen sanan frekvenssi vaikuttaa vahvasti.

Lähteet

Aineistolähteet

Helpuhe 2014 = Helsingin puhekielen pitkittäiskorpus (1970, 1990, 2010) [online puhekorpus]. Helsingin yliopiston suomen kielen, suomalais-ugrilaisten ja pohjoismaisten kielten ja kirjallisuuksien laitos, Kotimaisten kielten keskus ja Heikki Paunonen (2014). Kielipankki. http://urn.fi/urn:nbn:fi:lb-2014073041.

Kirjallisuuslähteet

Baayen, Harald R. – Endresen, Anna – Janda, Laura A. – Makarova, Anastasia – Nesset, Tore 2013: Making choices in Russian: pros and cons of statistical methods for rival forms. *Russian Linguistics* 37, 253–291. https://doi.org/10.1007/s11185-013-9118-6

Breiman, Leo – Friedman, J. H. – Olshen, R. A. – Stone, C. J. 1984: *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.

Bybee, Joan 2002: Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14, 261–290. https://doi.org/10.1017/S0954394502143018

Hakulinen, Lauri 2000: *Suomen kielen rakenne ja kehitys*. 5., muuttumaton painos. Helsingin yliopiston suomen kielen laitos, Helsinki.

Ho, Tin Kam 1995: *Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Vol. 1, 278–282.

Holm, Sture 1979: A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 65–70.

Hothorn, Torsten – Hornik, Kurt – Zeileis, Achim 2006: Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 15(3), 651–674. https://doi.org/10.1198/106186006X133933

Hothorn, Torsten – Hornik, Kurt – Zeileis, Achim 2021: *Party: A laboratory for recursive partitioning*. https://cran.r-project.org/web/packages/party/vignettes/party.pdf [luettu 30.3.2022.]

Kaakinen, Markus – Ellonen, Noora: Logistinen regressio. *Kvantitatiivisen tutkimuksen verkkokäsikirja*. Yhteiskuntatieteellinen tietoarkisto, Tampere. Saatavissa https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvanti/regressio/logistinen/ [luettu 3.1.2022].

Karlsson, Fred 2008: *Yleinen kielitiede*. Gaudeamus, Helsinki.

Kurki, Tommi 2005: *Yksilön ja ryhmän kielen reaaliaikainen muuttuminen. Kielenmuutosten seuraamisesta ja niiden tarkastelussa käytettävistä menetelmistä*. Suomalaisen Kirjallisuuden Seura, Helsinki.

Kuparinen, Olli 2021: *Muutoksen mekanismit. Kolmen aikapisteen reaaliaikatutkimus Helsingin puhekielestä*. Tampereen yliopiston väitöskirjat 428. Tampereen yliopisto, Tampere. https://urn.fi/URN:ISBN:978-952-03-1990-8

Kuparinen, Olli – Santaharju, Jenni – Leino, Unni – Mustanoja, Liisa – Peltonen, Jaakko 2022 (tulossa): Katomuotojen eteneminen yleiskielen *hd*-yhtymässä Helsingin puhekielessä. *Virittäjä* 126. https://doi.org/10.23982/vir.100585

Lehtimäki, Pekka 1983: Sekaidiolektien tutkimuksesta I. *Virittäjä* 87, 22–40.

Lehtinen, Tapani 2007: *Kielen vuosituhannet. Suomen kielen kehitys kantauralista varhaissuomeen*. Tietolipas 215. Suomalaisen Kirjallisuuden Seura, Helsinki.

Mantila, Harri 2004: Murre ja identiteetti. *Virittäjä* 108, 322–346.

Mielikäinen, Aila 1995: Morfologinen diffuusio. Morfologian asema äännehistoriallisessa tutkimuksessa. *Virittäjä* 99, 321–336.

Mustanoja, Liisa – O'Dell, Michael 2007: Suomen *d* ja *r* sosiofoneettisessa kentässä. *Virittäjä* 111, 56–67.

Mustanoja, Liisa 2011*: Idiolekti ja sen muuttuminen. Reaaliaikatutkimus Tampereen puhekielestä*.

Acta Universitatis Tamperensis 1605. Tampere University Press, Tampere. https://urn.fi/urn:isbn:978-951-44-8417-9

Nahkola, Kari 1987: *Yleisgeminaatio. Äänteenmuutoksen synty ja vaiheet kielisysteemissä erityisesti Tampereen seudun hämäläismurteiden kannalta*. Suomalaisen Kirjallisuuden Seuran Toimituksia 457. Suomalaisen Kirjallisuuden Seura, Helsinki.

Nuolijärvi, Pirkko 1986: *Kolmannen sukupolven kieli: Helsinkiin muuttaneiden suurten ikäluokkien eteläpohjalaisten ja pohjoissavolaisten kielellinen sopeutuminen*. Suomalaisen Kirjallisuuden Seura, Helsinki.

Paunonen, Heikki 2003: Suomen kielen morfologisista muutosmekanismeista. Lea Laitinen, Hanna Lappalainen, Päivi Markkola ja Johanna Vaattovaara (toim.): *Muotojen mieli. Kirjoituksia morfologiasta ja variaatiosta*, 187–248. Kieli 15. Helsingin yliopiston suomen kielen laitos, Helsinki.

Paunonen, Heikki 2009: Suomalaisen sosiolingvistiikan ja kielisosiologian näkymiä. *Virittäjä* 113, 557–570.

Priiki, Katri 2016: Puhutun suomen kielioppia ja yksilöllistä vaihtelua. Kvantitatiivinen tutkimus *hän*-pronominista Kaakkois-Satakunnan nykypuhekielessä. *Sananjalka* 58, 112–135. https://doi.org/10.30673/sja.86748

——— 2017: *Hän, se, tää vai toi? Vuorovaikutussosiolingvistinen tutkimus henkilöviittauksista Kaakkois-Satakunnan nykypuhekielessä*. Turun yliopiston julkaisuja 432. Turun yliopisto, Turku. https://urn.fi/URN:ISBN:978-951-29-6719-3

R Core Team 2021: *R: A language and environment for statistical computing*. 4.0.5. edn. Vienna, Austria: Foundation for Statistical Computing. https://www.R-project.org/.

Strasser, Helmut – Weber, Christian H. 1999: On the Asymptotic Theory of Permutation Statistics. *Mathematical Methods of Statistics* 8, 220–250.

Suihkonen, Paavo 1992: *Klusiilien vaihtelusuhteet Kala- ja Lestijokilaakson murteissa*. Suomalaisen Kirjallisuuden Seuran Toimituksia 577. Suomalaisen Kirjallisuuden Seura, Helsinki.

Wald, Abraham 1943: Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large. *Transactions of the American Mathematical Society* 54:3, 426–82. https://doi.org/10.2307/1990256

Wang, William S-Y 1969: Competing Changes as a Cause of Residue. *Language* 45, 9–25. https://doi.org/10.2307/411748

Väänänen, Milja 2016: *Subjektin ilmaiseminen yksikön ensimmäisessä persoonassa. Tutkimus suomen vanhoista murteista*. Turun yliopiston julkaisuja 430. Turun yliopisto, Turku. https://urn.fi/URN:ISBN:978-951-29-6664-6

Sini Knuutila, Olli Kuparinen, Jenni Santaharju, Liisa Mustanoja, Unni Leino, and Jaakko Peltonen: Why does the loss spread? Reasons for the diffusion of elision variants in hd *clusters in colloquial Finnish in Helsinki*

This article examines the reasons for elision variants of the *hd* cluster in standard Finnish (e.g. *ka**hd**eksan* 'eight') in the longitudinal corpus of colloquial Finnish spoken in Helsinki (e.g. *ka**h**eksan*). Plausible causes that are considered in this article are, for instance, word frequency and word formation, such as where the *hd* cluster is located. The cluster can appear in either vowel or consonant stems. Another plausible cause examined pertains to the meaning of the word, for example when comparing proper names to other words. The methodology used in this study includes logistic regression and the model of the decision tree.

The article shows that word frequency has the most impact on the probability of the *hd* cluster elision. Regarding word formation, the appearance of the *hd* cluster in a consonant stem after a long vowel, diphthong or weightless syllable makes the elision less probable. In terms of word meaning, proper names significantly show less the elision occurrences. Nonetheless, the effect of word frequency is the most relevant: if the word is not very suited for elision variants when it comes to word formation or meaning, its high frequency can still raise the probability of the elision. Respectively, in words where the prerequisites for the elision are favourable, its low frequency may keep the diffusion of the elision quite rare.

Sini Knuutila
sini.knuutila@tuni.fi
Tampereen yliopisto
https://orcid.org/0000-0001-6663-774X

Olli Kuparinen
Helsingin yliopisto
https://orcid.org/0000-0001-9468-7111

Jenni Santaharju
Helsingin yliopisto
https://orcid.org/0000-0002-7925-2715

Liisa Mustanoja
Tampereen yliopisto
https://orcid.org/0000-0003-1818-6926

Unni Leino
Tampereen yliopisto
https://orcid.org/0000-0003-3917-0026

Jaakko Peltonen
Tampereen yliopisto
https://orcid.org/0000-0003-3485-8585