ALEXANDRE NIKOLAEV JA JUSSI NIEMI

SUOMEN NOMINIEN TAIVUTUSJÄRJESTELMÄN PRODUKTIIVISUUDEN INDEKSEISTÄ

Miten taivutustyyppien produktiivisuus ilmenee kielen taivutusjärjestelmässä? Jos katsomme ongelmaa naiivin kielenpuhujan silmin (joka olisi kuitenkin tietoinen taivutustyyppien olemassaolosta), yksi toteamuksistamme lienee, että toisissa taivutustyypeissä on paljon sanoja kun taas toisissa niitä on vähän. Ehkä myös näkisimme taivutuksen jälkimmäisissä tyypeissä olevan jollakin tavalla mutkikkaampaa verrattuna edellisiin tyyppeihin. Mieleen voisivat tulla esimerkiksi lasten tekemät virheet tyyppiä **käsin* pro *käden*, joissa lapsi sijoittaa sanan väärään taivutustyyppiin.

Lingvisteinä pidämme yllä mainittuja toteamuksia taivutusjärjestelmän produktiivisuuden intuitiivisina indekseinä. Tämän kirjoituksen tarkoitus on kuitenkin nousta intuition yläpuolelle, kun operationaalistamme suomen nominien taivutusjärjestelmän produktiivisuutta korpuslingvistisin sekä laskennallisin menetelmin. Pyrimme samalla myös selvittämään, mitkä indeksit ovat relevantteja sellaisenaan ja mitkä ovat vuorovaikutuksessa keskenään.[1]

Morfologista produktiivisuutta on tutkittu sekä teoreettisesti (esim. Aronoff 1976; suomen kielessä esim. Penttilä 1963) että psykolingvistisesti (reaktioajan mittaamisesta ja silmän liikkeiden tutkimuksesta tilastollisten jakaumien analyysiin, ks. esim. Laudanna ja Burani 1995, Frauenfelder ja Shreuder 1992, Balota 1994, Baayen 1993; suomen kielessä esim. Niemi, Laine ja Tuominen 1994, Mäkisalo 2000, Järvikivi 2003, Nikolaev

[1] Haluamme kiittää kahta nimetöntä arvioijaa artikkelin kieliasua ja varsinkin sisältöä koskevista kommenteista ja ehdotuksista. Kiitämme myös tilastotieteen prof. Osmo Kolehmaista kommenteista.

ja Niemi 2006). Morfologisesti rikkaana kielenä suomi on hyvä tutkimuskohde morfologista produktiivisuutta ajatellen. Nimenomaan taivutusjärjestelmän rikkaus erottaa suomen indoeurooppalaisista kielistä, joissa morfologista produktiivisuutta on tutkittu perinteisesti joko sananjohdon tai verbien säännöllisen ja epäsäännöllisen taivutuksen osalta. Esimerkiksi Baayen (tulossa) raportoi tutkineensa 1600 englannin epäsäännöllisen ja säännöllisen verbin (suhde: 146/1454) tekstiyleisyyttä 18,5 miljoonan saneen korpuksessa. Laskelmien mukaan epäsäännöllisiä verbejä tavataan korpuksessa noin 2,5 kertaa useammin kuin säännöllisiä, vaikka jälkimmäisiä on otoksessa ollut noin 10 kertaa enemmän. Näin epäproduktiivisen kategorian suhteellisen korkea käyttöyleisyys suojaa Baayenin mukaan jäseniä siirtymisestä laajempaan produktiiviseen kategoriaan (tässä tapauksessa säännölliseen taivutukseen).

Epäproduktiiviset (kivettyneet) taivutustyypit ovat tyypillisesti suppeita ja niiden käyttöyleisyys on korkea, koska ne edustavat sanaston vanhinta kerrostumaa (englannin vanhimman kerrostuman korkeasta käyttöyleisyydestä ks. esim. Zipf 1947). Pagel, Atkinson ja Meade (2007) puoltavat tätä väitettä tarkastelemalla ja vertailemalla englannin lisäksi myös muita indoeurooppalaisia kieliä (venäjää, espanjaa ja kreikkaa). Tabak, Schreuder ja Baayen (2005: 9–11) toteavat hollannin verbien tarkastelun yhteydessä, että verbin etymologinen ikä korreloi sen taivutusparadigman kanssa siten, että vanhemman kerrostuman verbeillä on tilastollisesti suurempi todennäköisyys kuulua epäsäännölliseen taivutukseen.

Suomen taivutusjärjestelmä ei jakaudu kahtia säännölliseen ja epäsäännölliseen taivutukseen, vaan pikemmin voimme nähdä sen jatkumona produktiivisista täysin kivettyneisiin taivutustyyppeihin. Jatkumon toista päätä voi edustaa esimerkiksi produktiivinen taivutustyyppi *risti* kun taas toista päätä kivettynyt taivutustyyppi *käsi*.

Vertailemme seuraavaksi näiden produktiivisten ja kivettyneiden taivutustyyppien sanojen ikää ottamalla kummastakin taivutustyypistä kaksi 39 substantiivin otosta siten, että otokset vastaavat toisiaan sekä leksikaalisesti että fonologisesti.[2] Produktiivisen taivutustyypin otoksessa ainoastaan yksi sana (*mämmi*) kuuluu omaperäiseen sanastoon, kaikki muut ovat suhteellisen myöhäisiä lainoja joko ruotsista (22 sanaa) tai muinaisruotsista (6; lukuun ottamatta lainasanoja *lossi* (saksan kielestä) ja *laasti* (keskialasaksasta)). Kahdeksan sanan etymologia ei ole selvä. Kivettyneiden *i*-loppuisten taivutustyyppien sanoista vain yksi sana on lainattu muinaisruotsista (*tiili*), kaikki muut ovat joko vanhaa suomalais-ugrilaista (16) ja uralilaista (3) sanakerrostumaa tai hyvin vanhoja indoeurooppalaisia (4), balttilaisia (8) ja germaanisia (3) lainoja. Neljän sanan etymologia ei ole selvä.

[2] Sanojen oletettu ikä ja alkuperä perustuvat Nykysuomen etymologiseen sanakirjaan (Häkkinen 2004). Produktiivisen *risti*-taivutustyypin otos: *rahi, nänni, raspi, toti, paali, laasti, hormi, hauli, pässi, kloori, remmi, kurvi, mämmi, viiri, aasi, huivi, pahvi, selli, pihvi, malmi, koodi, kummi, peli, maali, tyyli, väri, kasvi, kahvi, pommi, pussi, kori, uuni, kuski, lasti, milli, konsti, kani, valssi, lossi*; kivettyneiden *i*-loppuisten taivutustyyppien otos: *ruuhi, jouhi, uuhi, mesi, loimi, pursi, orsi, tuohi, luomi, suoli, teeri, sääri, vuohi, riihi, liesi, nuoli, tiili, korsi, reisi, hirsi, köysi, liemi, vesi, käsi, saari, lumi, veri, susi, uni, kansi, sieni, lohi, kynsi, hiili, hiiri, veitsi, huuli, virsi, jousi*. Seuraavien muuttujien kohdalla otokset ovat verrannollisia: lemmataajuus (keskiarvo: 30,38 / 30,4), pintataajuus (4,84 / 4,82), fonologisten naapureiden lukumäärä (fonologisilla naapureilla tarkoitamme sanoja, jotka erottaa ainoastaan alkukirjain, engl. *neighborhood density*: 1,36 / 1,36), sanan pituus (4,89 / 4,74), sanan pareittaisten grafeemien keskimääräinen taajuus korpuksessa (engl. *bigram frequency*: 1531 / 1512).

▷

Kivettyneet *i*-loppuiset taivutustyypit (*käsi* jne.) edustavat siis odotetusti vanhempaa sanakerrostumaa kuin produktiiviset taivutustyypit (esim. *risti*). Erilaiset äänteenmuutokset ovat kohdistuneet kivettyneiden tyyppien taivutusparadigman eri kohtiin ja tuloksena on paradigma, joka sisältää paljon allomorfeja (esim. paradigma *käsi* sisältää allomorfit *käsi, käde, kät, käte* ja *käs*). Tämä tekee paradigmasta erittäin epäproduktiivisen. Toisin sanoen, jokin uusi *si*-loppuinen sana ei mene tähän paradigmaan, vaan mitä todennäköisemmin sana taivutetaan produktiivisen *risti*-paradigman mukaan. Koska kivettyneet taivutustyypit eivät enää tyypillisesti kasva, voimme olettaa, että niihin kuuluvia sanoja on vähän, mutta niiden käyttöyleisyys on suuri. Selityksenä siihen, miksi korkeampi käyttöyleisyys takaa niiden olemassaolon nykysuomessa, voimme käyttää kognitiivisen kielitieteen vakiintumisen (*entrenchment*) käsitettä (Braine ja Brooks 1995). Sen mukaan sanan taivuttaminen väärän paradigman mukaisesti riippuu siitä, kuinka usein lapsi kuulee sanan tietyssä konstruktiossa. Toisin sanoen, mitä tutumpi sana lapselle on, sitä harvemmin lapsi taivuttaa sitä väärin niissäkään konstruktioissa, joissa hän ei ole sitä vielä kuullut. Näin esimerkiksi pieni lapsi voi tehdä virheen **käsin* (pro *käden*), mutta vanhempi lapsi ei enää tee tätä virhettä, koska sana *käsi* on erittäin yleinen. Sana *heisi* on harvinainen, joten aikuinenkin natiivi suomen puhuja voi sanoa **heisin* (pro *heiden*; esim. *Itse istutin lumipalloheisin viime kesänä*... http://puutarha.net/index.asp?s=/keskustelu/keskustelu.asp?id=8914). Eli sanalla *käsi* on paljon korkeampi todennäköisyys pysyä kivettyneessä taivutustyypissä kuin sanalla *heisi*, joka ajan myötä voisi siirtyä produktiiviseen taivutustyyppiin *risti*.

Koska suomen taivutusjärjestelmä on jatkumo, jossa produktiivisten ja kivettyneiden tyyppien lisäksi pitäisi olla myös tyyppejä, joiden (epä)produktiivisuus ei ole absoluuttista, on mielenkiintoista testata, heijastuuko tämä jatkumo myös taivutustyypin koon ja käyttöyleisyyden vuorovaikutukseen. Hypoteesimme on, että taivutustyypin koko ja käyttöyleisyys korreloivat koko suomen taivutusjärjestelmässä. Testaamme tätä hypoteesia luvussa Taivutustyypin laajuus vs. käyttöyleisyys. Alaluvussa Hapaksien rooli ja liitteessä 2 laskemme myös taivutustyyppien produktiivisuuden asteet: sovellamme sananjohtotyyppien produktiivisuuden mittaamista varten luotua kaavaa taivutustyyppien produktiivisuuden mittaamiseen.

Luvussa Tavujen keskiarvo taivutustyypeissä pohdimme, korreloiko sanan pituus taivutustyypin (epä)produktiivisuuden kanssa ja luvussa Yhdyssanojen jakauma taivutustyypeittäin testaamme hypoteesia, jonka mukaan yhdyssanojen käyttöyleisyys tietyssä taivutustyypissä ei riipu sen produktiivisuudesta, sillä yhdyssanojen luokka on suomen kielessä erittäin produktiivinen ja avoin. Toisin sanoen vaikka kivettyneen taivutustyypin yksinkertaisia sanoja käytetäänkin keskimäärin useammin kuin produktiivisen, tämä ei koskisi näiden taivutustyyppien yhdyssanoja: kummassakin tyypissä yhdyssanojen keskimääräiset käyttöyleisyydet ovat oletettavasti samaa luokkaa.

Aineistona käytämme CD-Perussanakirjan (joka vuorostaan pohjautuu Suomen kielen perussanakirjaan, jäljempänä PS) tarjoamaa nominaalisten taivutustyyppien luokittelua (49 taivutustyyppiin)[3], johon kuuluu yhteensä 24 832 yksinkertaista lekseemiä ja 52 269

[3] Nomineilla tarkoitetaan substantiiveja, adjektiiveja, pronomineja ja numeraaleja. Pronomineista käytämme aineistossamme niitä, joilla on samanlainen paradigma muiden nominien kanssa (esim. *itse*, *eräs*, *kumpi*). Sanakirjojen tarjoama paradigmaklassifikaatio on ensisijaisesti tarkoitettu palvelemaan käytännöllisiä tarkoitusperiä (Karlsson 1983: 202). Laajana ja suhteellisen nykyaikaisena sanakirjana PS kuitenkin auttaa meitä

yhdyssanaa. Jälkimmäisiä tarkastelemme erikseen luvussa Yhdyssanojen jakauma taivutustyypeittäin. Näin ensimmäinen mitattu muuttuja on *leksikaalinen taajuus*, jolla tarkoitamme taivutustyypin laajuutta (sen jäsenten määrää), ja jonka laskemme CD-Perussanakirjasta. Esimerkiksi 5. taivutustyyppiin, jota paradigmaluokituksessa edustaa sana *risti*, kuuluu 4442 yksinkertaista lekseemiä, joten 5. taivutustypin leksikaalinen taajuus on 4442.

Toinen kvantitatiivinen muuttujamme on taivutustyyppien käyttöyleisyys eli *lemmataajuus*, jonka laskimme Kielipankin (www.csc.fi) 50 suomenkielisestä osakorpuksesta.[4] Tämän korpuksen saneiden määrä on 131 406 087. Laskimme kullekin lekseemille sen lemmataajuuden (tekstifrekvenssin) eli sen, kuinka monta kertaa kyseinen lekseemi esiintyy eri sananmuodossa mainitussa korpuksessa (esim. lekseemi *risti* esiintyy korpuksessa 6062 kertaa eri sananmuodossa: *risti, ristiä, risteissä,* jne., joten sen lemmataajuus on 6062). Emme kuitenkaan raportoi absoluuttista lemmataajuutta, vaan käytämme kielitieteellisen tradition mukaista lemmataajuutta per miljoona sanetta: näin laskettuna sanan *risti* lemmataajuus on 46,13. Teimme siis kaiken kaikkiaan noin 77 000 hakua yli 131 miljoonan saneen suuruisesta korpuksesta. Käytämme tätä korpusta, koska se on tällä hetkellä suomenkielisistä laajin, mikä takaa sovellettujen tilastollisten testien luotettavuuden. Toisaalta korpus ei ole tasapainotettu, eli se sisältää enimmäkseen lehtitekstejä eikä välttämättä aina heijasta sanojen yleisyyttä muissa tekstilajeissa tai puhutussa kielessä. Siksi esimerkiksi sanan *ponsi* lemmataajuus on suurempi kuin sanan *hiiri* (14,31 / 13,77), vaikka jälkimmäinen on puhutussa kielessä selvästi yleisempi kuin edellinen, joka on tyypillisempi sanomalehtikielelle.

TAIVUTUSTYYPPIEN LAAJUUS VS. KÄYTTÖYLEISYYS

Leksikaalinen taajuus on taivutustyypin ominaisuus (laskemme sen sanakirjasta) kun taas lemmataajuus on yksittäisen lekseemin ominaisuus (laskemme sen korpuksesta). Tämän vuoksi lemmataajuutta ei sellaisenaan voi käyttää taivutustyypin tarkastelussa, vaan on käytettävä sen sijasta jotakin sen johdannaista, kuten taivutustyyppiin kuuluvien lekseemien lemmataajuuksien mediaania.

Empiirisesti todetun Zipfin lain (Zipf 1935, 1949; Baayen 2001) mukaan sanan käyttöyleisyyden ja sen järjestysluvun välinen suhde on log–log-asteikolla lineaarinen. Sanan lemmataajuus on kääntäen verrannollinen sen järjestyslukuun: yleisin sana esiintyy kaksi kertaa niin usein kuin toiseksi yleisin sana, kolme kertaa niin usein kuin kolmanneksi yleisin sana ja niin edelleen. Helpottaaksemme muuttujien tulkintaa teimme niille logaritmimuunnoksen, jonka avulla Zipfin jakaumaa noudattavat muuttujat muuttuvat likimäärin lineaarisiksi. Ilman logaritmimuunnosta havainnot eivät olisi kovin informatiivisia: kuvaajassa

tarkastelemaan sanoja homogeenisina ryhminä, mikä on tärkeää esim. tilastollisten testien luotettavuuden kannalta. Tämän kirjoituksen tarkoitus ei ole esittää uutta paradigmaklassifikaatiota eikä arvostella PS:n tarjoamaa, vaikka jälkimmäinen perustuukin enimmäkseen tutkijoiden intuitioon.

[4] Aamulehti 1995, 1999; Demari 1995, 1997–2000; Hämeen Sanomat 1999–2000; Helsingin Sanomat 1995; Hyvinkään Sanomat 1994–1997; Iltalehti 1996; Kaleva 1998–1999; Kangasalan Sanomat; Karjalainen 1991–1995, 1997–1999; Karjalainen Määrittelemättömät; Keskisuomalainen 1999; Kustannusosakeyhtiö Otava 1993; Tekniikan Maailma 1995–1997; Turun Sanomat 1998–1999.

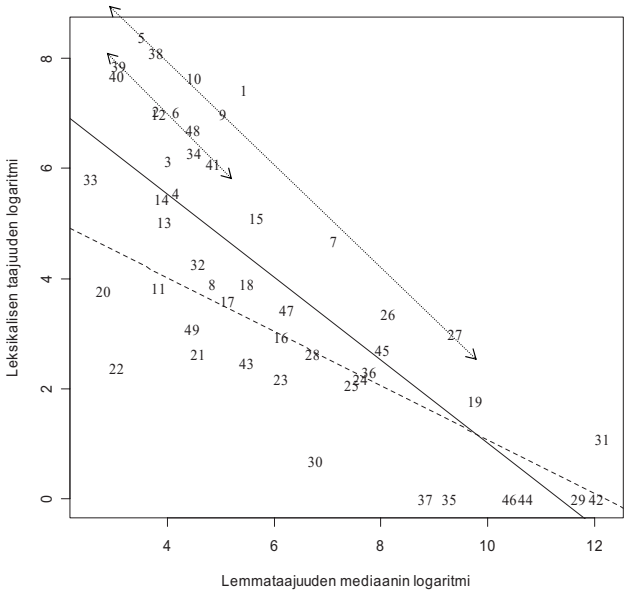
▷

ne kulkisivat koordinaattiakseleita pitkin, jolloin visuaaliseen tarkasteluun perustuvien päätelmien tekeminen olisi mahdotonta. Käytämmekin siis jäljempänä edellä mainittujen muuttujien logaritmisia arvoja. Kaikki muuttujat arvoineen on esitetty liitteessä 1.

Lemmataajuuden mediaani

Varhaisemmassa tutkimuksessamme (Nikolaev ja Niemi 2006) valitsimme mediaanin lemmataajuuden keskiluvuksi, koska se sopii hyvin tämän tyyppisiin vinoihin jakaumiin (*mediaani* on järjestetyn joukon keskimmäinen alkio). Tarkastelimme kahta kilpailijaparadigmaa 39. (*vastaus*) ja 41. (*vieras*) ja totesimme kieliin yleisesti sopivan havainnon, että produktiivinen taivutustyyppi (esim. *vastaus*) sisältää enemmän lekseemejä kuin epäproduktiivinen kilpailijansa (esim. *vieras*). Toisaalta produktiivisen taivutustyypin käyttöyleisyys on keskimäärin pienempi ja epäproduktiivisen vastaavasti suurempi.

Jos yleistämme yllä mainitun havainnon koskemaan koko nominaalista taivutusjärjestelmää, voimme odottaa muuttujien korreloivan keskenään: kun leksikaalinen taajuus kasvaa, lemmataajuuden mediaani laskee ja päinvastoin. Kuviossa 1 sovellamme näihin muuttujiin regressioanalyysia, jossa selitettäväksi muuttujaksi (*y*-akseli) valitsemme leksikaalisen taajuuden ja selittäjäksi (*x*-akseli) lemmataajuuden mediaanin. Emme kuitenkaan näe selitettävän ja selittävän muuttujan välillä suoraa syy–seuraus-suhdetta, muuttujilla on pikemmin yhteinen vaste — produktiivisuus, joka kuitenkin ilmenee muun muassa näiden muuttujien kautta (ks. perustelut jäljempänä esim. kuvio 2).



Kuvio 1. Taivutustyyppien leksikaalinen taajuus (logaritmimuunnos, *y*-akseli) vs. lemmataajuuden mediaani (logaritmimuunnos, *x*-akseli). Luvut 1–49 edustavat taivutustyyppejä (ks. liite 2); katkeamaton viiva on regressiosuora; katkoviiva on *bootstrap*-regressiosuora; kaksikärkiset nuolet kuvaavat kilpailijatyyppien 5, 7, 27 ja 39, 41 (epä)produktiivistumisen suuntaa.

Näiden kahden kilpailijaparadigman vertailun esimerkillä pyrimme vakuuttamaan lukijan siitä, että Nikolaevin ja Niemen (2006) tekemä toteamus pätee myös kaikkien muiden suomen nominaalisten taivutustyyppien kohdalla. Kuviosta 1 on helppo huomata muuttujien välinen lineaarinen riippuvuus: lemmataajuuden mediaanin kasvaessa leksikaalinen taajuus laskee. Muuttujien välillä on suhteellisen vahvaa negatiivista korrelaatiota (r=-0,77). Sovitettu (vapausasteilla korjattu) selitysaste R^2 on 0,59. Toisin sanoen taivutustyyppien käyttöyleisyys (lemmataajuuden mediaani; *x*-akseli) selittää merkitsevästi (p-arvo < 0,000 ***) noin 59 prosenttia leksikaalisen taajuuden (*y*-akseli) vaihtelusta.

Kuviosta 1 huomataan myös, että jäännösten (havainnon etäisyys regressiosuoralta) hajonta kasvaa systemaattisesti *x*-muuttujan arvojen muuttuessa: mitä pienempi lemmataajuuden mediaani, sitä suurempi jäännösten hajonta. Tilannetta kutsutaan tilastotieteessä *heteroskedastisuudeksi*. Voimme testata sitä Breusch-Paganin testin avulla (*Breusch-Pagan test for heteroskedasticity*). Testin p-arvot (0,06 ja 0,02; studentisoitu testi) puoltavat oletusta jäännösten heteroskedastisuudesta, jolla ei kuitenkaan ole haitallista vaikutusta regressiokertoimen arvoon. Sen sijaan sillä voi olla vaikutusta ennustamisen luotettavuuteen. Jos taivutustyypin lemmataajuuden mediaani on pieni, emme voi luotettavasti ennustaa kyseisen tyypin leksikaalista taajuutta.

Kuvion 1 regressiomalliin on sovellettu kahta suoraa: toinen on regressiomallin mukainen perinteinen pienimmän neliösumman (pns) regressiosuora (katkeamaton suora), kun taas katkoviivasuora kuvaa *bootstrap*-menetelmällä[5] tehtyä regressiota. Jos otetaan esimerkiksi kilpailijaparadigmapari 39 (*vastaus*) ja 41 (*vieras*), joista edellinen on selvästi produktiivinen ja jälkimmäinen epäproduktiivinen (ks. esim. Nikolaev 2002, Nikolaev ja Niemi 2006), ja piirretään näiden väliin viiva (kuvion 1 kaksikärkinen nuoli), huomataan sen olevan melkein paralleelinen pns-regressiosuoran kanssa. Samoin *i*-loppuiset taivutustyypit 5 (*risti*), 6 (*paperi*), 7 (*ovi*) ja 27 (*käsi*), joista ensimmäinen on produktiivisin ja jälkimmäinen vastaavasti epäproduktiivisin, ikään kuin laskeutuvat regressiosuoraa pitkin oikeaan alakulmaan. Jälkimmäisten taivutustyyppien järjestyksen jonossa näyttää määräävän tyyppien morfofonologinen kompleksisuus, joka kasvaa samalla kun regressiosuora laskee. Esimerkiksi paradigma *risti* on lähes agglutinatiivinen, kun taas paradigma *paperi* sisältää vaihtelua (*papereiden* vs. *paperien*; sisäisen vaihtelun vaikutuksesta tyyppien produktiivisuuteen ks. Nikolaev ja Niemi 2005), mikä tekee paradigmasta vähemmän produktiivisen. Epäproduktiivisessa paradigmassa *ovi* on *e*-vokaalivartalo, ja täysin kivettyneen *käsi*-paradigman allomorfit ovat melkein suppletiivisia. Paradigman tilalla voisi olla mikä tahansa kivettyneistä *i*-paradigmoista (taivutustyypit 23–31), sillä jälkimmäiset eivät ole enää lineaarisesti jakautuneita. Morfofonologinen kompleksisuus ei selitä niiden järjestystä kuvaajassa.

Produktiivisia taivutustyyppejä on siis vähän, ja ne keskittyvät kuvion 1 vasempaan yläkulmaan. Tätä toteamusta tukee myös *bootstrap*-menetelmällä tehty regressiosuora (katkoviivasuora kuviossa 1), joka pitää niitä tavallaan poikkeavina havaintoina. Sen sijaan produktiivisten taivutustyyppien jäsenten määrä (leksikaalinen taajuus) ylittää reilusti

[5] *Bootstrap*-menetelmä, toisin sanoen otoksen uusiokäyttömenetelmä, käyttää simulointia regressiomallissa. Otoksesta poimitaan pieniä otoksia palauttaen havainnot takaisin ja näihin perustuen saadaan testisuure. Toistoja voi olla tuhansia ja joissakin tilanteissa ne korvaavat epäparametrisia testejä. Menetelmää tarvitaan esim. silloin, kun normaalisuusoletus ei päde.

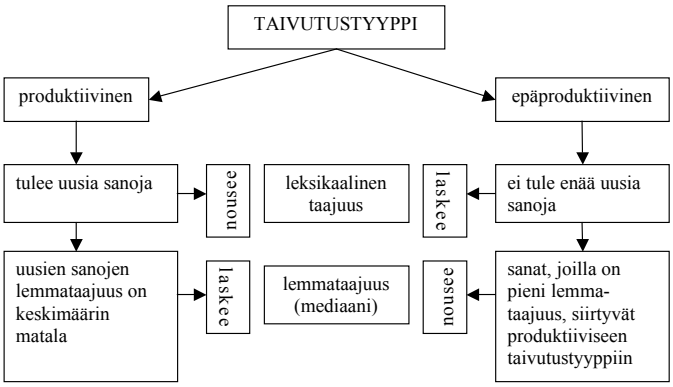
▷

epäproduktiivisten ja kivettyneiden taivutustyyppien jäsenten määrän. Esimerkiksi pelkästään jo tyypit 5 (*risti*), 38 (*nainen*), 39 (*vastaus*) ja 40 (*kalleus*) sisältävät yhteensä yli puolet kaikista lekseemeistä (12 520 / 24 832). Kärjessä olevaan 5. (*risti*) taivutustyyppiin sijoittuu valtaosa uudisnomineista (Karlsson 1983), ja vastaavasti tyyppeihin 38, 39 ja 40 sisältyy huomattava määrä produktiivisia johdoksia (Penttilä 1963). Regressiosuora siis näyttää taivutustyyppien produktiivistumisen ja epäproduktiivistumisen suunnan: pidämme mallia staattisena leikkauksena dynaamisesta systeemistä. Aiemmin tai myöhemmin tehty vastaava analyysi suomen yleiskielestä tuottaisi erittäin todennäköisesti toisenlaisen kuvion tai mallin.[6] Siksi käytämme nuolia ilmaisemaan kilpailijaparadigmojen välisiä suuntia.

Kuviossa 2 havainnollistamme leksikaalisen taajuuden ja lemmataajuuden (mediaanilla mitattuna) välistä vuorovaikutusta. Produktiivisissa tyypeissä jälkimmäinen pyrkii pienenemään samalla kun edellinen kasvaa. Epäproduktiivisissa tyypeissä tilanne on päinvastainen: tyypit menettävät jäseniä samalla periaatteella, jolla produktiiviset taivutustyypit kasvavat — »omasta hännästään» (hännällä tarkoitamme niiden sanojen joukkoa, joiden lemmataajuus on pieni). Otetaan esimerkin vuoksi epäproduktiivinen taivutustyyppi 7 (*ovi*) ja lajitellaan siihen kuuluvat sanat laskevaan järjestykseen niiden lemmataajuuden mukaan. Tuloksena huomataan epäproduktiivisille taivutustyypeille yleinen tendenssi: listan lopussa olevat sanat (joilla on pieni lemmataajuus) pyrkivät erkaantumaan joukosta. Tämä pyrkimys näkyy siinä, että näitä sanoja taivutetaan joko täysin tai rinnakkain kilpailijaparadigman mukaisesti. Esimerkiksi sanoista *haahti, palvi, hapsi* ja *helpi*, joiden lemmataajuudet ovat vastaavasti 1,23; 0,18; 0,07 ja 0,05, ei löytynyt korpuksesta yhtään *e*-vartaloista esiintymää. Sana *helpi* on toki palannut viime aikoina käyttöön, kun tutkijat huomasivat tämän kasvin olevan tehokas energialähde. Nyttemmin se sijoitetaan kuitenkin produktiiviseen paradigmaan *helpi : helpin* (Google-haussa (toukokuu 2007) tällaista taivutusta tavataan noin kaksi kertaa useammin kuin *helven*-sananmuotoa, joka on todennäköisesti kielenhuollon normeihin tutustuneiden toimittajien käsialaa).

Jos siis sana kuuluu epäproduktiiviseen taivutustyyppiin ja sen lemmataajuus on pieni, se on altis joko siirtymään produktiiviseen kilpailijaparadigmaan tai häviämään. Jälkimmäisestä kohtalosta sanaa voi tosin suojella sen morfologinen perhekoko (*family size*, ks. Nikolaev ja Niemi, käsikirjoitus), jolla mitataan sanan esiintymistä myös johdetuissa muodoissa ja yhdyssanoissa.

[6] Suomen kielen taivutusmorfologiahan on ja on ollut hienoisessa liikkeessä, joten meidän laskelmamme eivät ehkä hyvin kuvaisi esimerkiksi muutama kymmenen vuotta sitten vallinnutta tai erittäin todennäköisesti muutaman kymmenen vuoden päästä vallitsevaa järjestelmää (ks. esim. Anttila 1997).



Kuvio 2. Produktiivisuus yhteisenä vasteena leksikaalisen ja lemmataajuuden välisen yhteyden selityksessä.

Miten kunkin taivutustyypin lemmataajuuden mediaaniin vaikuttaa tyypin yleisimmän jäsenen esiintymien määrä (lemmataajuuden maksimiarvo)? Voisi nimittäin ajatella, että pienissä taivutustyypeissä on yleisimmän sanan taajuus suurempi kuin produktiivisemmissa tyypeissä, mikä selittäisi korkean käyttöyleisyyden mediaanin edellisissä ja matalan jälkimmäisissä. Jos siis selitämme leksikaalista taajuutta tyypin yleisimmän jäsenen lemmataajuudella, voisimme odottaa samanlaista negatiivista korrelaatiota kuin kuviossa 1. Näin ei kuitenkaan ole: mitä enemmän taivutustyypissä on jäseniä, sitä korkeampi on tyypin yleisimmän jäsenen lemmataajuus (korrelaatiokerroin on r=0,87). Maksimiarvo tunnuslukuna on sinällään hyvin herkkä satunnaisille vaihteluille, siksi regressiomallin selitysaste on pieni (12–14 % vaihtelusta), vaikka malli on edelleen merkitsevä (p-arvo: 0,008).

Vaikka produktiivisissa taivutustyypeissä tyypin yleisimmän jäsenen lemmataajuus on suurempi kuin epäproduktiivisissa tyypeissä, tyypin käyttöyleisyyden mediaani on kuitenkin pienempi, mikä selittyy juuri pienitaajuuksisilla sanoilla, joita on paljon enemmän produktiivisissa tyypeissä kuin epäproduktiivisissa (ks. perustelut edempänä ja taivutustyyppien jakaumat kuviosta 3 seuraavassa luvussa). Pienin mahdollinen sanan esiintymä korpuksessa on 1. Tällaisia sanoja on produktiivisessa tyypissä enemmän kuin epäproduktiivisessa, joten voisimme olettaa, että jos tiedämme niiden määrän tietyssä taivutustyypissä, voimme laskea tyypin produktiivisuuden asteen.

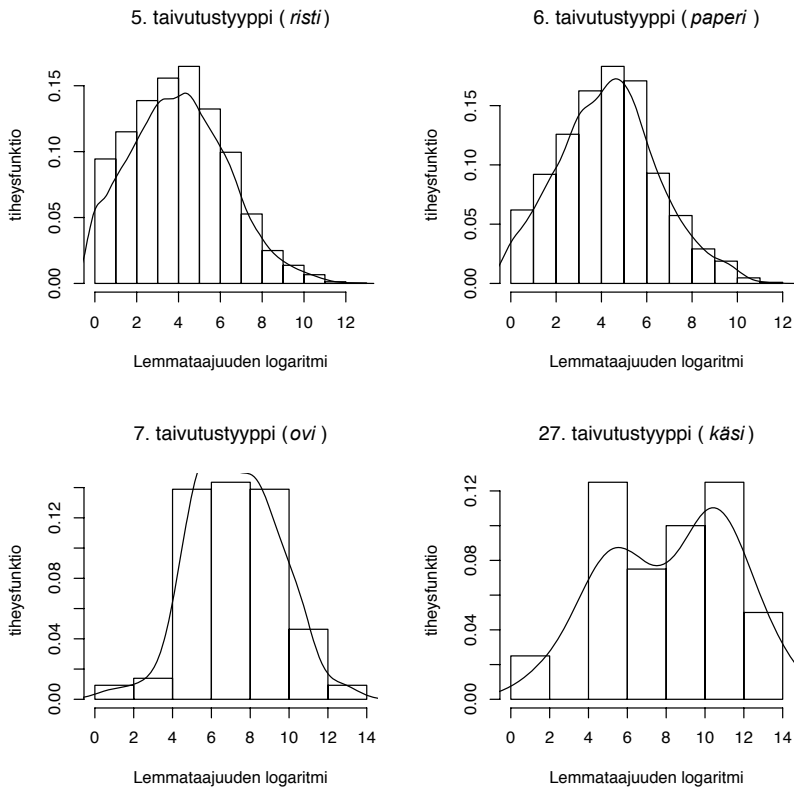
Hapaksien rooli

Hapaksi (*hapax legomenon*) tarkoittaa lekseemiä, jonka esiintymien määrä korpuksessa on tasan yksi. Harald Baayen käyttää tutkimuksissaan (esim. 1994, 2001, 2003) hapaksien määrää sananjohtotyyppien produktiivisuuden asteen mittaamisessa. Tämän ajattelun takana on hypoteesi kompleksisen sanan lemmataajuuden ja sanan morfien produktiivisuuden välisestä suhteesta: mitä suurempi kompleksisen sanan lemmataajuus on, sitä todennäköisemmin sana on tallennettu muistiin ja sitä vähäisempi rooli sanan ▷

morfologisilla komponenteilla on sen prosessoinnissa. Ja päinvastoin: jos kompleksisen sanan lemmataajuus on pieni, todennäköisesti sanansisäisellä morfologisella rakenteella on suurempi rooli sanan prosessoinnissa. (Bertram, Schreuder ja Baayen 2000; Baayen 2003: 241.)

Testaamme nyt Baayenin hypoteesia suomen taivutustyypeillä. Testaus on siksikin tarpeellista, että suomen taivutusmorfologia on esimerkiksi hollannin vastaavaa rikkaampaa ainakin kahdella tavalla: kilpailevia taivutusparadigmoja on useita ja paradigmat ovat laajoja (sija x luku) (suomen kilpailevien paradigmojen psykolingvistisestä tutkimuksesta ks. esim. Niemi 2006).

Epäproduktiivisen taivutustyypin 27 sanan *vuosi* lemmataajuus on suuri: 5812,69, ja sen morfologisesti monimutkainen paradigma on mitä todennäköisimmin memoroitu: *vuosi* : *vuoden* : *vuotta* ja niin edelleen. Voimmeko siis esimerkiksi odottaa lapsen taivuttavan sanaa *heisi* oikein — *heisi* : *heiden* : *heittä* ja niin edelleen, jos sen lemmataajuus on 0,05? Pienitaajuuksinen sanahan ei ole tyypillinen epäproduktiivisille taivutustyypeille. Sen sijaan produktiivisen 5. taivutustyypin sana *peski* on yhtä harvinainen (lemmataajuus: 0,05) ja sen paradigmaa *peski* : *peskin* : *peskiä* ei tarvinne memoroida erikseen. Kuviosta 3 näkyy, että produktiivisessa tyypissä jälkimmäisiä pienitaajuuksisia sanoja on enemmän kuin suuritaajuuksisia, mikä kääntyy peilikuvaksi epäproduktiivisissa taivutustyypeissä.



Kuvio 3. *i*-loppuisten taivutustyyppien jakaumat; lemmataajuus (*x*-akseli, logaritmimuunnos); tiheysfunktio (*y*-akseli).

Hapaksien määrä taivutustyypissä on sen käyttöyleisyyttä ilmaiseva muuttuja samalla tavalla kuin esimerkiksi lemmataajuuden mediaani. Kuten jälkimmäinenkin, se korreloi taivutustyypin leksikaalisen taajuuden kanssa: r=0,89 (p-arvo<0,000). Samoin kaksi kertaa korpuksessa esiintyvien sanojen määrä (*dis legomenon*) ja kolme kertaa esiintyvien sanojen määrä (*tris legomenon*) korreloi leksikaalisen taajuuden kanssa: kummallakin korrelaatiokerroin on 0,88 (p-arvo<0,000). Ainoa syy hapaksin valinnalle produktiivisuuden indeksiksi lienee se, että hapaksien esiintymien määrä on muita suurempi. Tämä selittyy Zipfin jakaumasta, jonka mukaan *m* kertaa esiintyneiden sanojen lukumäärä on kääntäen verrannollinen sanan lemmataajuuteen (*m*), toisin sanoen suuri määrä sanakirjan sanoja esiintyy korpuksessa harvoin ja vielä suurempi määrä ei esiinny korpuksessa lainkaan (*out-of-vocabulary*, OOV: noin 10 % PS:n nomineista (2541/24 832) ja melkein kolmasosa (30,3 %) on lemmataajuudeltaan korkeintaan 10 (*m=10*; ks. taulukkoa 1).

Taulukko 1. *m* kertaa esiintyneiden lekseemien taajuus (m: 1–10).

m	[0]	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
V_m	2541	1061	799	594	483	421	412	356	320	289	254

Myös neljä ja viisi kertaa korpuksessa esiintyvien sanojen määrä taivutustyypissä korreloi tyypin leksikaalisen taajuuden kanssa: r=0,86 (p-arvo<0,000), eli tendenssi on selvä — samalla kun lemmataajuus (*m*) kasvaa, korrelaatiokerroin ja esiintymien määrä (V_m) pienenevät, mikä vuorostaan noudattaa Zipfin lakia.

Baayen luokittelee produktiivisuuden asteen hapaksiehdolliseksi (*hapax-conditioned degree of productivity*, *P**) ja kategoriaehdolliseksi (*category-conditioned degree of productivity*, *P*). Nimittäjänä edellisessä (*P**) hän käyttää hapaksien määrää koko korpuksessa: (V_1, *N*) ja jälkimmäisessä (*P*) taas tietyn morfologisen kategorian (tässä tapauksessa taivutustyypin) sanojen määrää (N_i). Kummankin osoittajana on tiettyyn morfologiseen kategoriaan kuuluvien hapaksien määrä (V_1, *N*, *i*):

$$P^* = (V_1, N, i) / (V_1, N); P = (V_1, N, i) / N_i$$

*P** on todennäköisyys sille, että sana kuuluu tiettyyn kategoriaan, kun jo tiedetään, että se esiintyy aineistossa täsmälleen kerran. Leksikaalinen taajuus on pohjimmiltaan yhtäpitävä sen todennäköisyyden kanssa, että (aineistosta sattumanvaraisesti poimittu) sana kuuluu tiettyyn taivutustyyppiin. Myös hapaksiehdollinen produktiivisuusaste on määritelmänsä mukaan todennäköisyys sille, että sana kuuluu tiettyyn taivutustyyppiin, mutta nyt mukana on lisärajoite, että sanan on esiinnyttävä aineistossa täsmälleen kerran. Toisin sanoen — jos tulkitaan harvoin esiintyvät sanat produktiivisesti taivutettaviksi vastakohtana taivutukseltaan kiteytyneille — hapaksiehdollinen produktiivisuusaste kertoo, kuinka suuri osuus »produktiivisesti» taivutettavista sanoista kuuluu tähän tyyppiin.

Vastaavasti produktiivisuusaste *P* on todennäköisyys sille, että sana esiintyy korpuksessa täsmälleen kerran, kun tiedetään sen kuuluvaan tiettyyn kategoriaan. Se siis kertoo,

▷

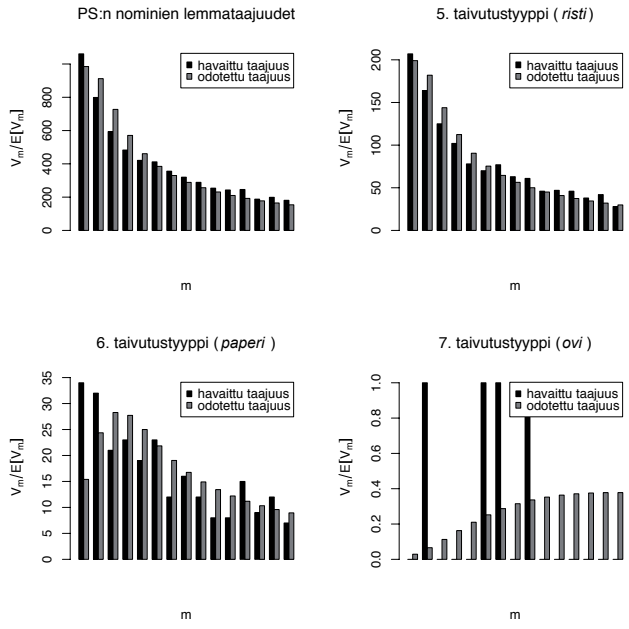
kuinka suuri osuus taivutustyyppiin kuuluvista sanoista taivutetaan »produktiivisesti». Näin tämä muuttuja on riippumattomampi taivutustyypin laajuudesta.[7]

Suomen kielen morfologian produktiivisuuden tutkimuksessa kaavaa *P* ovat aiemmin soveltaneet Juhani Järvikivi (2003: 28–30) ja Jukka Mäkisalo (2000: 14). Järvikivi laski sanomalehti Karjalaisen korpuksen perusteella (34,5 miljoonaa sanaa) produktiivisuuden asteen kymmenelle sananjohtotyypille. Mäkisalo sovelsi tätä kaavaa tutkiessaan yhdyssanojen produktiivisuutta.

Näin sovellettuna produktivisuuden mitta jättää kuitenkin selittämättä kaikki epäproduktiiviset ja kivettyneet taivutustyypit, sillä useimmiten niissä ei tavata hapaksia ollenkaan (ks. esim. kuvio 3). Lisäksi laajatkin korpukset ovat kuitenkin aina rajallisia. Siksi tarvitsemme hapaksin estimoitua arvoa, jota Baayen käyttää myöhemmissä tutkimuksissaan (esim. 2003) produktiivisuuden asteen laskemisessa:

$$P^* = \mathrm{E}[(V_1, N, i)] / \mathrm{E}[(V_1, N)]; \; P = \mathrm{E}[(V_1, N, i)] / N_i$$

Kuviossa 4 vertaamme havaittuja ja estimoituja[8] taajuuksia koko aineistossa (PS) ja taivutustyypeissä 5, 6 ja 7. Pylväät ovat järjestyksessä m=1, m=2 ja niin edelleen siten, että kukin pylväs esittää m kertaa esiintyneiden sanojen lukumäärää (esim. $V_{m=1}$ = niiden sanojen lukumäärä, joiden lemmataajuus on 1 (*hapax legomenon*) ja vastaavasti $V_{m=2}$ = *dis legomenon* jne.). Havainnollisuuden vuoksi esitämme kussakin kuvaajassa 15 ensimmäistä pylväsparia (max (m) = 15).



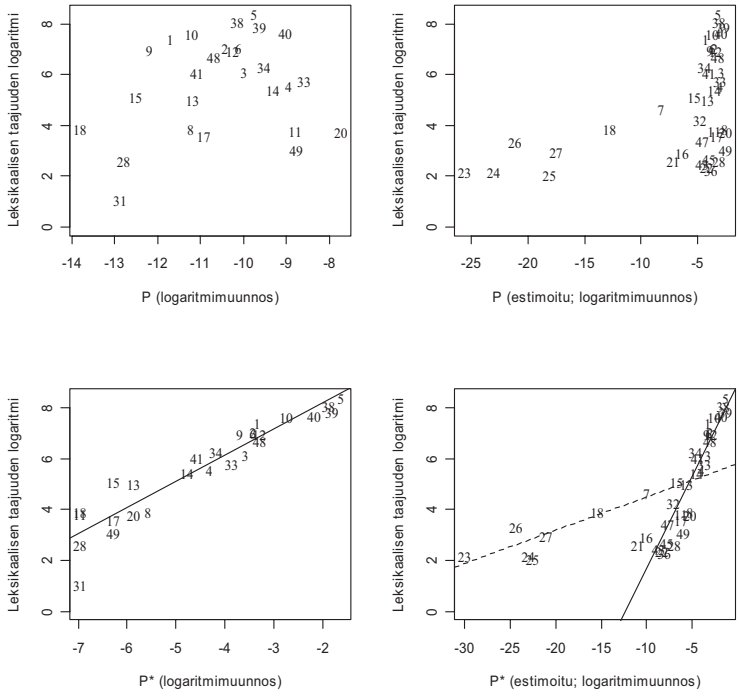
Kuvio 4. *m* kertaa esiintyneiden sanojen havaittu vs. odotettu taajuus (*x*-akseli, m: 1–15); *y*-akseli: esiintymien määrä.

[7] Käytämme tässä artikkelin arvioijan ehdottamaa produktiivisuusasteiden (*P* ja *P**) selitystä.

[8] Estimoidut taajuudet laskimme *finite Zipf-Mandelbrot* (fZM) LNRE (*large number of rare events*)-mallin avulla (yksityiskohdista ja matemaattisesta perustelusta ks. Evert 2004).

Kuviosta 4 huomataan, että produktiivisen taivutustyypin 5 (*risti*) jakauma on heijastuma koko sanaston (PS) jakaumasta, sillä kummatkin noudattavat Zipfin jakaumaa ja estimoidut taajuudet ovat likimain samanlaisia kuin havaitut taajuudet. Sen sijaan (sisäisen vaihtelun takia) vähemmän produktiivisessa tyypissä 6 (*paperi*) odotettu hapaksien määrä on pienempi kuin havaittu ja epäproduktiivisessa tyypissä 7 (*ovi*) ei ole yhtään hapaksia, mutta estimoitu määrä on kuitenkin nollaa suurempi: 0,03, mistä näkyy toisaalta estimoinnin hyöty (estimoitujen parametrien avulla voimme laskea produktivisuuden asteen myös epäproduktiivisille ja kivettyneille tyypeille), ja toisaalta estimoinnin vaarallisuus pienellä aineistolla (mitä pienempi aineisto, sitä suurempi voi olla estimoinnin virhe).

Liitteessä 2 on laskettu yllä mainitut produktiivisuuden asteet (*P*, P*) kaikille suomen nominien taivutustyypeille, joissa tavataan hapakseja. Koska monissa tyypeissä ei esiinny hapaksia, laskimme kaikille tyypeille produktiivisuuden asteen myös estimoiduilla parametreilla (lukuun ottamatta tyyppejä, joiden leksikaalinen taajuus on pienempi kuin 8 (pienin määrä, josta tässä aineistossa voi estimoida arvoja): 19. (*suo*), 31. (*kaksi*), 30. (*veitsi*), 29. (*lapsi*), 35. (*lämmin*), 37. (*vasen*), 42. (*mies*), 44. (*kevät*), 46. (*tuhat*).[9]



Kuvio 5. Taivutustyyppien leksikaalinen taajuus (*y*-akseli, logaritmimuunnos) vs. produktiivisuuden aste (*P*, *P**, estimoitu *P*, estimoitu *P** , *x*-akseli, logaritmimuunnos).

[9] Parametrit estimoitiin Sichelin Gauss-Poisson LNRE -mallin (The Generalized Inverse Gauss-Poisson (GIGP) LNRE model of Sichel (1971)) avulla, joka kuuluu samaan luokkaan (FZM) LNRE -mallin kanssa (ks. kuvio 4) ja on rakennettu Zipf-Mandelbrotin lain perusteella (ks. esim. Baayen 2001: 82–93).

▷

Kuviossa 5 selitämme taivutustyyppien leksikaalista taajuutta produktiivisuuden asteilla (ks. liite 2). Oikeanpuoleiset kuvaajat antavat laajemman perspektiivin, sillä selitämme näissä leksikaalista taajuutta myös sellaisissa taivutustyypeissä, joissa ei ole yhtään hapaksia. Kuvaaja on sikäli mielenkiintoinen, että havainnot erottuvat tavallaan kahdeksi regressiomalliksi: epäproduktiiviset *i*-loppuiset (katkoviivasuora) vs. kaikki muut taivutustyypit (katkeamaton suora). Molemmat regressiomallit ovat merkitseviä.[10] Estimoitu produktiivisuuden aste *P** selittää siis kivettyneiden *i*-loppuisten tyyppien jakaumaa paremmin kuin lemmataajuuden mediaani (ks. kuvio 1). Kategoriaehdollinen produktiivisuuden aste *P* (*x*-akseli, vasen ylimmäinen kuvaaja kuviossa 5) ei selitä taivutustyyppien leksikaalista taajuutta, ja vastaavasti estimoiduilla parametreilla laskettu (*x*-akseli, oikea ylimmäinen kuvaaja) muistuttaa hapaksiehdollista *P**-astetta.

Baayenin mukaan (2003: 241) kategoriaehdollinen *P* perustuu suoraan morfologiseen kategoriaan (tässä tapauksessa taivutustyyppiin) ja indeksoi tyypin produktiivisuutta ja potentiaalia epäsystemaattisista tekijöistä riippumatta, kun taas hapaksiehdollinen *P** on herkkä erilaisille epäsystemaattisille tekijöille. Eri tavalla lasketut taivutustyyppien produktiivisuuden asteet antavat kukin erilaisia arvoja samoille taivutustyypeille ja ovat näin ristiriitaisia tulkinnan kannalta. Tämä teoreettisesti kylläkin mielenkiintoinen tulos vaatii tarkastelua psykolingvististen kokeiden avulla, joita aiommekin jatkossa suorittaa.

TAVUJEN KESKIARVO TAIVUTUSTYYPEISSÄ

Suomen kielessä tavujen määrä sekä tavujen raskaus (mora-arvo) ohjaavat sanan taivutusta, esimerkiksi sanat *maa*, *kala* ja *satama* kuuluvat eri taivutustyyppeihin (ks. myös tavun raskauden vaikutuksesta nominien taivutukseen Nikolaev ja Niemi 2005). Voisimme odottaa, että tavujen määrä sanaa kohti kussakin taivutustyypissä ilmaisee tyypin produktiivisuutta ja myös korreloi muiden produktiivisuuden indeksien kanssa, kuten tyyppien leksikaalisen ja lemmataajuuden kanssa. Laskeaksemme tavujen keskiarvon kussakin tyypissä jaoimme kaikki sanat (24 832) tavuihin.[11]

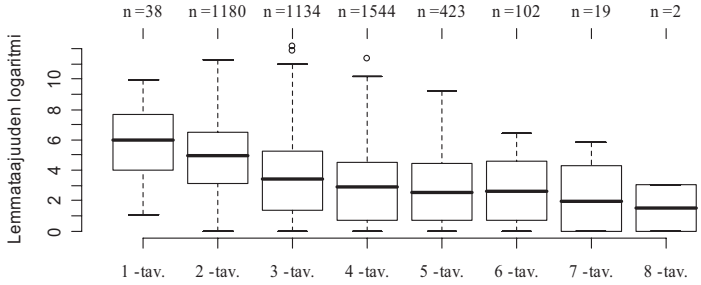
Yksi Zipfin laeista (1935) toteaa, että kielen yleisimmät sanat ovat suhteellisen lyhyitä. Testatkaamme nyt tätä väitettä vaikkapa 5. (*risti*) taivutustyypin esimerkillä jakamalla sen kaikki lemmat (4 442) kahdeksaan eri ryhmään tavumäärän mukaan ja laskemalla lemmataajuuden kullekin sanalle kussakin ryhmässä. Kuviosta 6 näkyy, että ryhmät noudattavat Zipfin lakia: mitä vähemmän sanoissa on tavuja, sitä korkeampi on lemmataajuuden mediaani (1-tavuisten ryhmään kuuluvat sellaiset sanat, kuin *golf*, *pop*, *jazz*, *rap*, *zen* jne., fonologisesti sanat ovat kuitenkin kaksitavuisia: *golfi, poppi* jne.).[12]

---

[10] *i*-loppuiset taivutustyypit: R^2=0,66; p-arvo=0,015. Muut taivutustyypit: R^2=0,85; p-arvo<0,000.

[11] Jaon tavuihin tilasimme Lingsoftilta, jossa se tehtiin Finhyph- ja Finhyphpro-ohjelmien avulla. Koska kyseiset ohjelmat on luotu tekstinkäsittelyä varten, ne jättivät tavuttamatta osan alku- ja lopputavuista (esim. *ora-va*). Tavutimme listan loppuun käyttäen omia perl-skriptejämme.

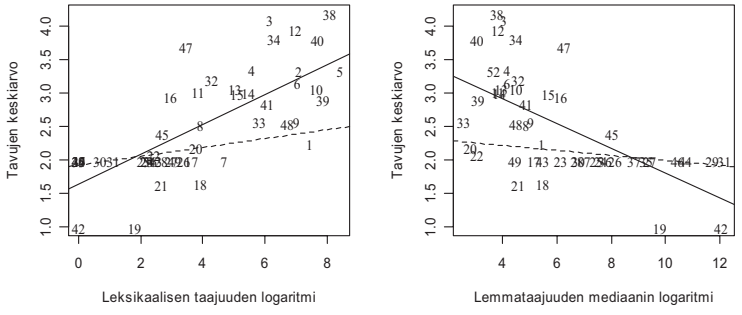
[12] Kruskal-Wallisin testin mukaan (epäparametrinen vastine yksisuuntaiselle varianssianalyysille) ryhmien välillä on tilastollisesti merkitsevä ero: testisuure=473,13, vapausasteet=7, *p*-arvo<0,000 (5 % merkitsevyystasolla seuraavat ryhmät eroavat keskenään: 1-3, 1-4, 1-5, 1-6, 2-3, 2-4, 2-6, 2-7, 3-6, 3-7, 4-5, 5-6, 5-7).



Kuvio 6. 5. (*risti*) taivutustyypin 8 eri ryhmän (tavuluvun perusteella) lemmataajuus (logaritmimuunnos).

Voisiko tämän tuloksen, joka koskee yhden taivutustyypin sisäistä jakaumaa, yleistää taivutustyyppien väliseen jakaumaan? Vastaus on kyllä, jos käytämme taivutustyypin käyttöyleisyyden indeksinä mediaania (ks. kuvion 7 oikeanpuoleista kuvaajaa). Mitä produktiivisempi tyyppi on, sitä pienempi on sen käyttöyleisyyden (lemmataajuuden) mediaani (ks. perustelut edempänä, kuvio 1 ja kuvio 3). Tämä vuorostaan selittää, miksi produktiivisten taivutustyyppien sanat ovat keskimäärin pitempiä kuin epäproduktiivisten taivutustyyppien. Kuten näemme kuviosta 7, tavujen keskiarvo korreloi positiivisesti leksikaalisen taajuuden kanssa (r=0,73; p-arvo=0,000) ja negatiivisesti lemmataajuuden mediaanin kanssa (r=-0,62; p-arvo=0,000).

Kuvion 7 oikeanpuoleinen kuvaaja ei kuitenkaan ole yksiselitteisesti merkitsevä. Jos nimittäin lasketaan muuttujille osittaiskorrelaatiot (kahden muuttujan välinen korrelaatio, kun kolmannen muuttujan vaikutus on poistettu molemmasta muuttujasta), niin edellinen korrelaatio on 0,50 ja jälkimmäinen -0,13. Eli tavun keskimääräinen pituus ja lemmataajuuden mediaani eivät korreloi silloin, kun muiden muuttujien vaikutus on eliminoitu; näiden yhteinen vaste on leksikaalinen taajuus. Näin sanan keskimääräinen pituus taivutustyypin sisällä ei ainakaan suoranaisesti korreloi taivutustyypin käyttöyleisyyden kanssa.



Kuvio 7. Vasemmanpuoleinen kuvaaja: Tavujen keskiarvo taivutustyypissä (y-akseli) vs. leksikaalinen taajuus (logaritmimuunnos; x-akseli). Oikeanpuoleinen kuvaaja: Tavujen keskiarvo taivutustyypissä (y-akseli) vs. lemmataajuus (logaritmimuunnos; x-akseli).

▷

Kummankin kuvaajan kärjessä on joukko taivutustyyppejä, joiden tavujen keskiarvo on suurempi kuin 3,5: 3. (*valtio*), 12. (*kulkija*), 34. (*onneton*), 38. (*nainen*), 40 (*kalleus*), 47 (*kuollut*). Kaikki nämä taivutustyypit ovat johdosluonteisia (Penttilä 1963) ja niiden monimorfeemisuus selittää korkeamman tavujen keskiarvon verrattuna muihin produktiivisiin tyyppeihin.

Tavujen keskiarvo taivutustyypissä (kuvio 7) auttaa meitä jakamaan kuvion 1 produktiivisten tyyppien joukon kahtia produktiivisuuden suhteen. Monimorfeemisia taivutustyyppejä 3, 12, 34, 38, 40 ja 47 (ks. yllä) voimme siis pitää produktiivisina suppeassa mielessä, koska ne sisältävät produktiivisia ja erittäin produktiivisia johtimia (Penttilä 1963: 280, 281, 295, 298, 302, 306), mutta niiden produktiivisuus ei kuitenkaan tarkoita paradigman avoimuutta. Esimerkiksi automalli Volkswagen Phaeton taipuu *phaeton* : *phaetonin* eikä **phaettoman* (vrt. **Ostin eilen phaettoman*). Meillä siis pitää olla leksikaalista (maailman)tietoa (eikä pelkästään morfologista), jotta voisimme valita taivutusmuotojen *phaettoman* ja *phaetonin* välillä. Näin taivutustyypit 34. (*onneton*) ja 5 (*risti*) ovat kumpikin produktiivisia, mutta edellisen produktiivisuuden taustalla on produktiivinen johdin.

Toinen mielenkiintoinen ryhmä erottuu, jos valitsemme tavujen keskiarvon ylärajaksi arvon 2. Kahdeksallatoista taivutustyypillä tavujen keskiarvo on tasan 2 ja neljällä alle 2. Näiden yhteinen leksikaalinen taajuus on 359 (!, vrt. muiden 27 taivutustyypin yht. leksikaaliseen taajuuteen 24 473). Erittäin pieni leksikaalinen taajuus vihjaa taivutustyypin kaksitavuisuuden olevan kivettyneiden taivutustyyppien ominaisuus (esim. kaikki kivettyneet *i*-loppuiset taivutustyypit eli luokat 23–31 ovatkin kaksitavuisia). Kuitenkin tämän ryhmän vaikutus on suuri, ja se näkyy *bootstrap*-regressiomallissa (katkoviivasuorat kuviossa 7), jossa leksikaalinen taajuus ja lemmataajuus eivät selitä enää tavujen keskiarvoa taivutustyypeissä.

Suomen vapaiden morfeemien (perusmuotojen) suosituin kanoninen rakenne on kuitenkin kaksitavuinen (ks. esim. Karlsson 1983: 217). Koska kivettyneet taivutustyypit edustavat leksikon vanhaa kerrostumaa, ei ole sattumaa, että juuri ne ovat kaksitavuisia. Toisaalta myös produktiiviset taivutustyypit suosivat kaksitavuisuutta, sillä esimerkiksi ruotsin ja varsinkin englannin (joka on nykyään tärkein lainanantajakieli) yksitavut tyyppiä CVC lainataan suomeen kaksitavuisina ja *i*-loppuisina (5. taivutustyypin kaksitavuisten ryhmän — ks. kuvio 6 — leksikaalinen taajuus ja lemmataajuus ovat korkeita). Karlsson (mts. 218) mainitsee myös affektipitoisten sanojen suosivan kaksitavuisuutta: *eka* (taivutustyyppi 9; lemmataajuus 21,13), *toka* (10; 0,21), *telkku* (1; 0,14), *nekru* (1; 0,35) jne. Sanojen »produktiivisuus» näkyy mm. siinä, että ne eivät ole kvalitatiivisen astevaihtelun alaisia (*ekan, tokan*).

Näin voimme todeta, että tavujen keskiarvo erottaa kaksi ryhmää taivutustyyppejä: i) kivettyneet taivutustyypit, jotka ovat staattisia tavumäärän suhteen ja prototyyppisesti kaksitavuisia, ja ii) johdosluonteiset produktiiviset taivutustyypit, jotka ovat monimorfeemisia ja siksi pyrkivät kohti monitavuisuutta. Muut (epä)produktiiviset taivutustyypit eivät erotu yhtä merkitsevästi tavujen keskiarvon suhteen.

YHDYSSANOJEN JAKAUMA TAIVUTUSTYYPEITTÄIN

Yhdyssanoja on aineistossamme noin kaksi kertaa enemmän kuin yksinkertaisia sanoja (52 269/24 832). Tämä suhde ei kuitenkaan säily kaikissa taivutustyypeissä samanlaisena. Jäljempänä vertailemme yhdyssanojen ja yksinkertaisten sanojen suhdetta taivutustyypeittäin.

Taulukossa 2 vertailemme yksinkertaisten ja yhdyssanojen määrää yhdeksässä yleisimmässä taivutustyypissä (taulukon vasen puoli) ja yhdeksässä kivettyneessä *i*-loppuisessa tyypissä (oikea puoli). Kummassakin ryhmässä yhdyssanoja on enemmän kuin yksinkertaisia (Wilcoxonin testin[13] p-arvo=0,004). Kuitenkin suhde on näissä kahdessa ryhmässä erilainen: kivettyneille taivutustyypeille on ominaista ylivoimainen yhdyssanojen dominointi, mikä on seurausta kyseisten tyyppien jakaumasta (kuvio 3). Epäproduktiivisissa tyypeissä on enemmän suuritaajuuksisia sanoja, kun taas produktiivisissa tyypeissä pienitaajuuksisia. Näin kivettyneiden tyyppien prototyyppisenä edustajana voisi olla mikä tahansa suuritaajuuksinen lekseemi, kuten *lapsi*, ja produktiivisten vastaavasti mikä tahansa pienitaajuuksinen lekseemi, kuten *rapsi* (vrt. näiden potentiaalia esiintyä yhdyssanojen loppuosana: *lapsi* 38, *rapsi* 0). Tässä kuitenkin kannattaa toistaa, että aineistossamme on vain niitä yhdyssanoja, jotka löytyvät PS:sta; korpuksesta löytyy lisää enemmän tai vähemmän leksikaalistuneita yhdyssanoja, kuten esimerkiksi sana *rehurapsi*, ja erilaisia *lapsi*-loppuisia yhdyssanoja on korpuksessa 620 (*sijoituslapsi* (lemmataajuus 0,21), *muslimilapsi* (0,08), *aviolapsi* (0,05) jne.; vrt. PS:ssa 38).

Taulukko 2. 9 yleisintä taivutustyyppiä vs. kivettyneet *i*-loppuiset taivutustyypit (leks. t. = leksikaalinen taajuus; yhd.l.t. = yhdyssanojen leksikaalinen taajuus ko. taivutustyypissä).

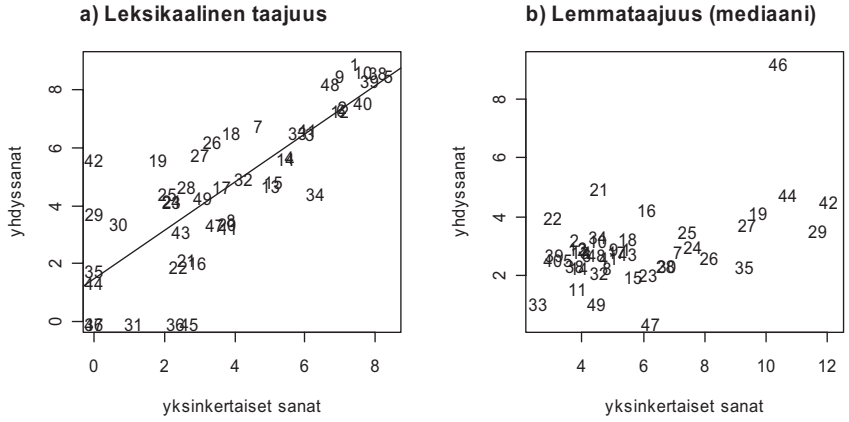
taivutustyyppi	leks.t.	yhd.l.t.	taivutustyyppi	leks.t.	yhd.l.t.
5. (*risti*)	4442	5183	26. (*pieni*)	29	454
38. (*nainen*)	3287	5643	27. (*käsi*)	20	309
39. (*vastaus*)	2622	4333	28. (*kynsi*)	14	87
40. (*kalleus*)	2169	2046	23. (*moni*)	9	54
10. (*koira*)	2092	5883	24. (*uni*)	9	70
1. (*valo*)	1667	7940	25. (*toimi*)	8	76
2. (*palvelu*)	1172	1765	31. (*kaksi*)	3	0
6. (*paperi*)	1142	1583	30. (*veitsi*)	2	26
9. (*kala*)	1090	5305	29. (*lapsi*)	1	38

Kuviossa 8a vertaamme yksinkertaisten ja yhdyssanojen leksikaalisen taajuuden suhdetta taivutustyypeissä. Havainnot ovat jakautuneet siten, että vasemmassa alakulmassa niiden residuaalit (jäännökset) ovat kaikkein suurimpia (heteroskedastisuuden Breusch-

[13] Riippuvien otosten t-testin epäparametrinen vastine (Wilcoxon Matched Pairs Signed Rank Test , ks. esim. Hollander ja Wolfe 1973).

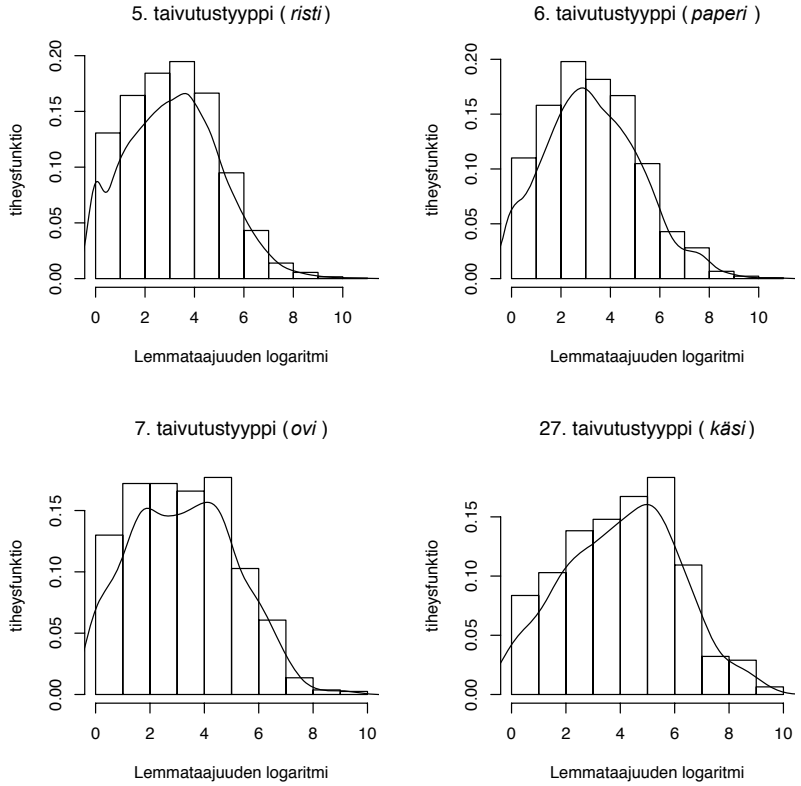
▷

Paganin testin p-arvo=0,004). Siitä huolimatta riippuvuus on erittäin merkitsevä (sovitettu selitysaste: 0,65; F(1,47)=90,31; p-arvo<0,000). Heteroskedastisuuden selittää toisaalta aineistomme luonne: sanakirjassa ei esimerkiksi luetella yhdysnumeraaleja, siksi tyypeissä 31 (*kaksi*) ja 45 (*kahdeksas*) ei ole yhdyssanoja; ja toisaalta on itsestään selvää, että yhden sanan taivutustyypissä 42 (*mies*) on enemmän yhdyssanoja (273) kuin esimerkiksi taivutustyypissä 36 (*sisin*), jossa yhdyssanan loppuosa olisi superlatiivissa (*alin, ylin, lähin* jne.). Produktiivisissa taivutustyypeissä taas yhdyssanojen ja yksinkertaisten sanojen suhde on lineaarinen (kuvaajan oikea yläkulma).



Kuvio 8. a) Taivutustyyppien leksikaalinen taajuus (logaritmimuunnos): yhdyssanat (*y*-akseli) vs. yksinkertaiset sanat (*x*-akseli). b) Taivutustyyppien lemmataajuuden mediaani (logaritmimuunnos): yhdyssanat (*y*-akseli) vs. yksinkertaiset sanat (*x*-akseli). Luvut 1–49 edustavat taivutustyyppejä (ks. liitteet).

Vastaako kussakin tyypissä yhdyssanojen käyttöyleisyys (lemmataajuus) yksinkertaisten sanojen käyttöyleisyyttä (lemmataajuutta)? Kuvio 8b ei osoita selvää lineaarista riippuvuutta tyyppien lemmataajuuden mediaanilla mitattuna: missään taivutustyypissä yhdyssanojen käyttöyleisyyden mediaani ei riipu yksinkertaisten sanojen käyttöyleisyyden mediaanista. Yhdyssanojen käyttöyleisyys ei siis riipu suoraan taivutustyypin produktiivisuudesta. Näin on luultavasti siksi, että yhdyssanat edustavat sekä suomessa että muissa kielissä, joissa sananmuodostuskeinona on johto, erittäin avointa ja produktiivista luokkaa (ks. esim. Dressler 2006). Kivettyneiden taivutustyyppien yhdyssanat eivät tarvitse korkeaa lemmataajuutta samalla tavalla kuin yksinkertaiset sanat selvitäkseen elossa. Havainnollistamme tätä kuvion 9 avulla, joka esittää *i*-loppuisten taivutustyyppien yhdyssanojen jakaumaa. Kuviosta huomataan, että epäproduktiivisten tyyppien 7 (*ovi*) ja 27 (*käsi*) jakaumat ovat verrannollisia produktiivisten 5 (*risti*) ja 6 (*paperi*) tyyppien jakaumiin. Sen sijaan vastaavat yksinkertaisten sanojen jakaumat eroavat selvästi produktiivisissa ja epäproduktiivisissa tyypeissä (kuvio 3).



Kuvio 9. *i*-loppuisten yhdyssanojen taivutustyyppien jakaumat (*x*-akseli: lemmataajuuden logaritmimuunnos; *y*-akseli: tiheysfunktio.

PÄÄTELMIÄ

Tärkeä kysymys tässä tutkimuksessa on myös kielen ajallinen dynaamisuus, sillä produktiivisuuden tutkimuksen kautta voimme tehdä ennustuksia siitä, mihin suuntaan suomen taivutusjärjestelmä muuttunee, kun samasta lekseemistä voi kilpailla useampikin paradigma. Yksittäisen sanan taajuudella (käyttöyleisyydellä) on tärkeä rooli kielen muutoksessa (esim. Zipf 1949; Bybee ja Hopper 2001). Epäproduktiiviset taivutustyypit eivät tyypillisesti enää kasva, vaan — päinvastoin — ne menettävät jäseniä, sillä pienitaajuuksinen sana on altis joko siirtymään produktiiviseen paradigmaan tai jäämään pois käytöstä. Hiljattain on löydetty (Lieberman, Michel, Jackson, Tang ja Nowak 2007) matemaattinen riippuvuus sanan taajuuden ja sen ajan välillä, jolloin sana siirtyy epäproduktiivisesta paradigmasta produktiiviseen. Tutkijat tarkastelivat englannin epäsäännöllisten verbien taivutusta 1200 vuoden ajalta ja totesivat seuraavan säännönmukaisuuden: epäsäännöllisen verbin siirtymiseen säännölliseen paradigmaan kulunut aika on verbin käyttöyleisyyden neliöjuuri. Toisin sanoen, jos toisen epäsäännöllisen verbin lemmataajuus on 100 kertaa pienempi kuin toisen, se siirtyy säännölliseen taivutukseen 10 kertaa nopeammin. Tulosten pätevyys muissa kielissä kaipaa vielä tarkastelua. Kuitenkin voimme havain- ▷

nollistaa tätä kaavaa kahden 7. taivutustyypin sanan esimerkillä: jos oletetaan, että sana *helpi* (lemmataajuus 0,05) produktiivistuu (siirtyy kokonaan 5. taivutustyyppiin *helpi* : *helpin*), niin sana *torvi* (lemmataajuus 4,44) hypoteettisesti produktiivistuu kymmenen kertaa hitaammin (*torvi* : *torvin*).

Artikkelin alussa kuviteltu naiivi puhuja totesi, että laajan taivutustyypin paradigma on helpompi (lue: *produktiivisempi*) kuin suppeahko taivutustyyppi, jonka taivutusparadigma tuntuu vaikeammalta. Tämä asiantila näkyy esimerkiksi lasten ja vieraskielisten puhujien tekemissä virheissä. Laskemamme hapaksiehdollinen produktivisuuden aste (sekä estimoitu että tavallinen *P**) asettaa taivutustyypit melkein samanlaiseen järjestykseen produktiivisuuden suhteen kuin leksikaalinen taajuus (ks. liitteet 1 ja 2). Sen sijaan kategoriaehdollinen produktiivisuuden aste *P* ei suoranaisesti korreloi taivutustyypin laajuuden kanssa. Siksi Baayen (esim. 1994: 466) pitää sitä relevantimpana produktiivisuuden indeksinä.

Suoritamme jatkossa psykolingvistisiä kokeita tarkistaaksemme eri produktiivisuusasteiden pätevyyttä tyyppien produktiivisuuden selittäjinä. Voimme kuitenkin alustavasti (karkeasti) testata sitä kolmen kilpailijaparin avulla: *nalle* vs. *hame*, *vastaus* vs. *kalleus* ja *risti* vs. *paperi*.

Ensimmäisen parin 8. taivutustyyppi *nalle* on produktiivinen verrattuna tyyppiin 48 *hame* (vrt. lasten tekemät virheet taivutustyypissä *hame*: Räisänen 1975; Karlsson 1983; Niemi ja Niemi 1987), vaikka *nalle*-tyypin leksikaalinen taajuus on pienempi kuin *hame*-tyypin (50 vs. 828).

Toisen parin ehdokkaat ovat pikemmin komplementaarisessa distribuutiossa keskenään: taivutustyypin 40 *kalleus* nomineista 84,6 % (1836/2169) on *UUs*-loppuisia, kun taas taivutustyypissä 39 *vastaus* vain yksi nomini on *uus*-loppuinen: *makuus* (rinn. *makaus*). Muut kuin *UUs*-loppuiset *kalleus*-tyypin nominit (15,6 %) ovat herkempiä siirtymään kilpailijatyyppiin *vastaus*. Näin esimerkiksi eräs radiokuuluttaja sanoi vahingossa muutaman kerran: »Aamuhartauksessa puhuu...» (Maisa Martin, suullisesti Kielitieteen päivillä 2007 Oulussa). Koska kyseessä on radio-ohjelman nimi, joka on tavallaan ikonisempi kuin ominaisuuden nimi, sana on siirtynyt sille produktiivisempaan paradigmaan *vastaus*. Samalla tavalla sana *Varkaus* kaupungin nimenä on ikonisempi kuin *varkaus* (rikos) ja on siksi joskus altis siirtymään tyyppiin *vastaus*: esimerkiksi *Varkauksen paikallissää* (http://www.tietotori.fi/Varkaus/Varkaus%20Julkiset/ I000A46B0). Koska ei ole havaittavissa päinvastaista tendenssiä taivuttaa 39. tyyppiin kuuluvia sanoja *kalleus*-tyypin mukaan, voimme todeta 39. taivutustyypin *vastaus* olevan produktiivisempi kuin 40 *kalleus*. Näiden tyyppien leksikaalinen taajuus on melkein samanlainen: 2622 (*vastaus*) ja 2169 (*kalleus*).

Kolmannen parin ehdokkaista taivutustyyppi *risti* 5 on ylivoimaisesti produktiivisin, ja *paperi* 6 on vähemmän produktiivinen sisäisen vaihtelun takia. Tämä myös näkyy niiden leksikaalisessa taajuudessa: 4442 vs. 1142.

Seuraavassa taulukossa vertaamme liitteessä 2 esitettyjä produktiivisuusasteita yllä käsiteltyjen kilpailijaparien avulla:

Taulukko 3. Kilpailijaparien *nalle* vs. *hame*, *vastaus* vs. *kalleus* ja *risti* vs. *paperi* vertailu produktiivisuusasteiden avulla.

Taivutustyyppi		***P***	**Taivutustyyppi**		***P****
40	(kalleus)	0,00012	5	(risti)	0,1951
39	(vastaus)	6,53E-05	39	(vastaus)	0,16494
5	(risti)	5,80E-05	40	(kalleus)	0,11216
6	(paperi)	4,03E-05	48	(hame)	0,03676
48	(hame)	2,26E-05	6	(paperi)	0,03205
8	(nalle)	1,31E-05	8	(nalle)	0,00377
Taivutustyyppi		***P* (estimoitu)**	**Taivutustyyppi**		***P**** **(estimoitu)**
8	(nalle)	0,07215	5	(risti)	0,24323
39	(vastaus)	0,06474	39	(vastaus)	0,2409
40	(kalleus)	0,05157	40	(kalleus)	0,15873
48	(hame)	0,0396	48	(hame)	0,04653
5	(risti)	0,03859	6	(paperi)	0,04011
6	(paperi)	0,02475	8	(nalle)	0,00512

Taulukosta näkyy, että ainut produktiivisuusaste, joka asetti kaikki kolme kilpailijaparia ennustamaamme keskinäiseen järjestykseen, on kategoriaehdollinen estimoitu *P*. Liitteestä 2 huomataan, että sama *P* (estimoitu) asettaa ensimmäiselle sijalle taivutustyypin 49 *askel*, mikä ainakin intuitiivisesti tuntuu oudolta. Tässä pienessä taivutustyypissä (leksikaalinen taajuus on 22) tavataan yksi hapaksi: sana *säen*/*säkene*. Produktiivisuusasteiden takana on ajatus, että jokainen hapaksi on tavallaan neologismi, uudismuodoste. Koska neologismi on useimmiten produktiivisen taivutustyypin ominaisuus, ratkaisee neologismien (hapaksien) määrä viime kädessä, kuinka produktiivisena voimme pitää tiettyä taivutustyyppiä. Sana *säen*/*säkene* on pikemmin arkaismi kuin neologismi ja näin se kuvaa päinvastaista tendenssiä: arkaismien ja neologismien käyttöyleisyys voi olla yhtä pieni, mutta edellisessä tapauksessa se todennäköisesti laskee, kun taas jälkimmäisessä nousee ajan myötä.

Tavujen keskiarvo taivutustyypissä ei suoranaisesti kerro tyypin produktiivisuudesta: muuttujan kärjessä olevat taivutustyypit ovat johdosluonteisia, sillä monimorfeemisuutensa ansiosta ne ovat keskimäärin muita pitempiä. Yhdyssanojen luokka on suomen kielessä erittäin produktiivinen, siksi niiden käyttöyleisyysjakauma produktiivisessa tyypissä on verrannollinen jakaumaan epäproduktiivisessa tyypissä.

Olemme todenneet tässä tutkimuksessa kaksi tendenssiä, jotka ovat ominaisia myös muille kielille (ks. esim. Baayen 1994: 467): (i) produktiivisten taivutustyyppien keskimääräinen käyttöyleisyys on pieni ja epäproduktiivisten suuri, ja (ii) produktiivinen taivutustyyppi on suurempi kuin epäproduktiivinen. Laajan aineiston perusteella havaitaan tilastollisesti erittäin merkitsevä korrelaatio taivutustyypin käyttöyleisyyden ja laajuuden välillä. Kumpikin näistä kahdesta muuttujasta on omalla tavallaan tyypin produktiivisuuden indeksi, jota produktiivisuusasteen rinnalla voimme käyttää selittäjänä, kun analysoimme psykolingvististen kokeiden tuloksia.

▷

Toisessa tutkimuksessamme (Niemi, Nikolaev ja Hugdahl tulossa) vertaamme 7-, 10- ja 14-vuotiaiden lasten epäsanatesteissä saatuja tuloksia niiden koehenkilöiden vastauksiin, joille on diagnosoitu sukutaustainen dysfasia (*specific (familial) language impairment*). Regressioanalyysin avulla selitämme siinä oikein taivutettujen epäsanojen jakaumaa taivutustyyppien lemmataajuudella ja leksikaalisella taajuudella. Jälkimmäinen muuttuja selittää tuloksia keskimäärin paremmin (merkitsevämmin) kuin edellinen; kuitenkin tendenssi on molemmissa analyyseissa sama: morfologista kehitystä (järjestelmän stabilisointia) tapahtuu normaaleilla lapsilla vielä melko myöhään, tässä tutkimuksessa 7–14 vuoden ikäisinä. Sen sijaan puhujilla, joilla on sukutaustainen dysfasia, paradigman valitseminen ei korreloi produktiivisuuden kanssa. Ko. tutkimus antaa siis lisätukea tässä artikkelissa esitettyjen tilastomatemaattisten tulosten pätevyydelle.

LÄHTEET

ANTTILA, ARTO 1997: *Variation in Finnish phonology and morphology*. Painamaton väitöskirja. Yleisen kielitieteen laitos, Stanfordin yliopisto.

ARONOFF, MARK 1976: *Word formation in generative grammar.* Cambridge, MA: MIT Press.

BAAYEN, R. HARALD 1993: On frequency, transparency and productivity. – Geert E. Booij & Jaap van Marle (toim.), *Yearbook of morphology 1992* s. 181–208. Dordrecht: Kluwer.

——— 1994: Productivity in language production. – *Language and Cognitive Processes* 9 s. 447–469.

——— 2001: *Word frequency distributions*. Dordrecht: Kluwer.

——— 2003: Probabilistic approaches to morphology. – Rens Bod, Jennifer B. Hay & Stefanie Jannedy (toim.), *Probabilistic linguistics* s. 229–287. Cambridge, MA: The MIT Press.

——— (tulossa): Corpus linguistics in morphology: Morphological productivity. – Anke Lüdeling, Merja Kytö & Tony McEnery (toim.), *Handbook of corpus linguistics*. Berlin: De Gruyter.

BALOTA, DAVID A. 1994: Visual word recognition: The journey from features to meaning. – Morton Ann Gernsbacher (toim.), *Handbook of psycholinguistics* s. 303–348. San Diego, CA: Academic Press.

BERTRAM, RAYMOND – SCHREUDER, ROBERT – BAAYEN, R. HARALD 2000: The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. – *Journal of experimental psychology: Learning, memory, & cognition* 26 s. 489–511.

BRAINE, MARTIN D. S. – BROOKS, PATRICIA J. 1995: Verb argument structure and the problem of avoiding an overgeneral grammar. – Michael Tomasello & William Edward Merriman (toim.), *Beyond names for things: Young children's acquisition of verbs* s. 353–376. Hillsdale, NJ: Erlbaum.

BYBEE, JOAN – HOPPER, PAUL J. (toim.) 2001: *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamins.

CD-Perussanakirja. Kotimaisten kielten tutkimuskeskuksen julkaisuja 94. Helsinki: Ko-

timaisten kielten tutkimuskeskus 1997.

Dressler, Wolfgang U. 2006: Compound types. – Gary Libben & Gonia Jarema (toim.), *The representation and processing of compound words* s. 23–44. New York: Oxford University Press.

Evert, Stefan 2004: A simple LNRE model for random character sequences. – *Proceedings of JADT* 2004 s. 411–422.

Frauenfelder, Ulli – Schreuder, Robert 1992: Constraining psycholinguistic models of morphological processing and representation: The role of productivity. – Geert Booij & Jaap van Marle (toim.), *Yearbook of morphology* 1991 s. 165–183. Dordrecht: Kluwer.

Hollander, Myles – Wolfe, Douglas A. 1973: *Nonparametric statistical methods*. New York: John Wiley & Sons.

Häkkinen, Kaisa 2004: *Nykysuomen etymologinen sanakirja*. Helsinki: WSOY.

Järvikivi, Juhani 2003: *Allomorphy and morphological salience in the mental lexicon*. Joensuu: Joensuun yliopisto.

Karlsson, Fred 1983: *Suomen kielen äänne- ja muotorakenne*. Helsinki: WSOY.

Laudanna, Alessandro – Burani, Christina 1995: Distributional properties of derivational affixes: Implications for processing. – Laurie Beth Feldman (toim.), *Morphological aspects of language processing: Cross-linguistic perspectives* s. 345–364. Hillsdale, NJ: Erlbaum.

Lieberman, Erez – Michel, Jean-Babtiste – Jackson, Joe – Tang, Tina – Nowak, Martin A. 2007: Quantifying the evolutionary dynamics of language. – *Nature* 449 s. 713–716.

Mäkisalo, Jukka 2000: *Grammar and experimental evidence in Finnish compounds*. Joensuu: Joensuun yliopisto.

Niemi, Jussi 2006: Paradigm competition: An experimental note on Finnish verbs. – Michael Suominen, Antti Arppe, Anu Airola, Orvokki Heinämäki, Matti Miestamo, Urho Määttä, Jussi Niemi, Kari Pitkänen & Kaius Sinnemäki (toim.), A Man of Measure : A Festschrift in honour of Fred Karlsson on his 60th Birthday. Helsinki: The Linguistic Association of Finland. *SKY Journal of linguistic*s, Special Issue of Vol. 19 s. 227–235.

Niemi, Jussi – Laine, Matti – Tuominen, Juhani 1994: Cognitive morphology in Finnish: Foundations of a new model. – *Language and Cognitive Processes* 9 s. 423–446.

Niemi, Jussi – Niemi, Sinikka 1987: Acquisition of inflectional marking: A case study of Finnish. – *Nordic journal of linguistics* 10 s. 59–89.

Niemi, Jussi – Nikolaev, Alexandre – Hugdahl, K. (tulossa): Impaired performance in attention to phonological input, dysfunctional lexical »frequency counter/s» and abnormal grammatical morphology: A possible causal chain? – J. Zlatev, M. Johansson Falck, C. Lundmark & M. Andrén (toim.)*, Studies in Language and Cognition*. Newcastle: Cambridge Scholars Publishing.

Nikolaev, Alexandre 2002: Eräiden suomen taivutustyyppien produktiivisuudesta. – *Puhe ja kieli* 22 s. 113–124.

Nikolaev, Alexandre – Niemi, Jussi 2005: Suomen nominien taivutuksesta: rytmi-, sivupaino- ja agglutinaatiohypoteesien testausta. – *Virittäjä* 109 s. 530–553.

▷

——— 2006: Nominien paradigmaattistuminen suomessa. Millä rakenteellisilla ehdoilla kielenkäyttäjät sitovat potentiaaliset nominit taivutusluokkiin? – *Virittäjä* 110 s. 46–69.
——— (käsikirjoitus): Suomen yhdyssanojen produktiivisuudesta.
Pagel, Mark – Atkinson, Quentin. D. – Meade, Andrew. 2007: Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. – *Nature* 449 s. 717–721.
Penttilä, Aarni 1963: *Suomen kielioppi*. Toinen, tarkistettu painos. Helsinki: WSOY.
PS = *Suomen kielen perussanakirja*. Helsinki: Valtion painatuskeskus ja Kotimaisten kielten tutkimuskeskus 1990–1994.
Räisänen, Alpo 1975: Havaintoja lastenkielestä. – *Virittäjä* 79 s. 251–266.
Sichel, H. S. 1971: On a family of discrete distributions particularly suited to represent long-tailed frequency data. – *Proceedings of the third symposium on mathematical statistics* s. 51-97.
Tabak, Wieke M. – Schreuder, Robert – Baayen, R. Harald 2005: Lexical statistics and lexical processing: Semantic density, information complexity, sex, and irregularity in Dutch. – Marga Reis & Stephan Kepser (toim.), *Linguistic Evidence* s. 529–555. Berlin: Mouton de Gruyer.
Zipf, G. K. 1935 *The Psycho-Biology of Language*. Boston: Houghton Mifflin.
——— 1947: Prehistoric 'cultural strata' in the evolution of Germanic: The case of Gothic. – *Modern language notes* 62 s. 522–530.
——— 1949: *Human behavior and the principle of the least effort. An introduction to human ecology*. New York: Hafner.

LIITTEET

Liite 1. Artikkelissa käytetyt muuttujat logaritmimuunnoksineen. Kukin rivi antaa käyttämämme ao. taivutustyypin arvot (leksikaalinen taajuus, lemmataajuuden mediaani ja maksimiarvo, tavujen keskiarvo ja hapaksien määrä).

		yksinkertaiset sanat									yhdyssanat			
taivutustyyppi	mallisana	leksikaalinen taajuus	log(leksikaalinen taajuus)	mediaani(lemmataajuus)	log(mediaani(lemmataajuus))	max(lemmataajuus)	log(max(lemmataajuus))	tavujen keskiarvo	hapaksien määrä	log(hapaksien määrä)	leksikaalinen taajuus	log(leksikaalinen taajuus)	mediaani(lemmataajuus)	log(mediaani(lemmataajuus))
1	(valo)	1667	7,42	227	5,42	129408	11,77	2,25	38	3,64	7940	8,98	19	2,94
2	(palvelu)	1172	7,07	44	3,78	71396	11,18	3,34	34	3,53	1765	7,48	26	3,26
3	(valtio)	467	6,15	56	4,03	78260	11,27	4,11	29	3,37	688	6,53	19,5	2,97
4	(laatikko)	263	5,57	64	4,16	12729	9,45	3,36	14	2,64	312	5,74	17,5	2,86
5	(risti)	4442	8,4	34	3,53	189601	12,15	3,34	207	5,33	5183	8,55	13	2,56
6	(paperi)	1142	7,04	64	4,16	67773	11,12	3,18	34	3,53	1583	7,37	16	2,77
7	(ovi)	108	4,68	1240	7,12	439549	12,99	2	0	0	919	6,82	17	2,83
8	(nalle)	50	3,91	126,5	4,84	142084	11,86	2,54	4	1,39	36	3,58	10	2,3
9	(kala)	1090	6,99	154	5,04	1190160	13,99	2,59	26	3,26	5305	8,58	19	2,94
10	(koira)	2092	7,65	90	4,5	284956	12,56	3,08	68	4,22	5883	8,68	25	3,22
11	(omena)	46	3,83	46,5	3,84	2072	7,64	3,04	1	0	27	3,3	5	1,61
12	(kulkija)	1080	6,98	46,5	3,84	196831	12,19	3,96	39	3,66	1546	7,34	19	2,94
13	(katiska)	153	5,03	52	3,95	83800	11,34	3,07	3	1,1	119	4,78	17	2,83
14	(solakka)	231	5,44	49	3,89	13769	9,53	3,02	9	2,2	292	5,68	10	2,3
15	(korkea)	167	5,12	287	5,66	87978	11,38	3	2	0,69	136	4,91	7,5	2,01
16	(vanhempi)	19	2,94	457	6,12	40479	10,61	2,95	0	0	8	2,08	72,5	4,28
17	(vapaa)	37	3,61	170	5,14	27420	10,22	2	2	0,69	111	4,71	17	2,83
18	(maa)	50	3,91	242,5	5,49	229990	12,35	1,66	1	0	735	6,6	27	3,3
19	(suo)	6	1,79	17337	9,76	152945	11,94	1	0	0	278	5,63	64,5	4,17
20	(filee)	44	3,78	16,5	2,8	1459	7,29	2,18	3	1,1	31	3,43	0	0
21	(rosé)	14	2,64	97,5	4,58	1652	7,41	1,64	0	0	9	2,2	152	5,02
22	(parfait)	11	2,4	21	3,04	7346	8,9	2,09	0	0	7	1,95	55	4,01
23	(moni)	9	2,2	463	6,14	136230	11,82	2	0	0	68	4,22	8	2,08
24	(uni)	9	2,2	2018	7,61	14177	9,56	2	0	0	71	4,26	21	3,04
25	(toimi)	8	2,08	1716	7,45	2096	7,65	2	0	0	89	4,49	34	3,53
26	(pieni)	29	3,37	3408	8,13	247339	12,42	2	0	0	523	6,26	14	2,64
27	(käsi)	20	3	12004	9,39	763823	13,55	2	0	0	347	5,85	43	3,76
28	(kynsi)	14	2,64	834,5	6,73	15311	9,64	2	1	0	111	4,71	11	2,4
29	(lapsi)	1	0	119530	11,69	119530	11,69	2	0	0	43	3,76	36	3,58
30	(veitsi)	2	0,69	868,5	6,77	1523	7,33	2	0	0	31	3,43	11	2,4
31	(kaksi)	3	1,1	189032	12,15	205478	12,23	2	1	0	0	0	0	0
32	(sisar)	71	4,26	97	4,57	57733	10,96	3,21	0	0	150	5,01	8,5	2,14
33	(kytkin)	333	5,81	13	2,56	27758	10,23	2,59	22	3,09	725	6,59	3	1,1
34	(onneton)	541	6,29	90	4,5	24373	10,1	3,83	16	2,77	87	4,47	28	3,33
35	(lämmin)	1	0	10734	9,28	10734	9,28	2	0	0	6	1,79	10,5	2,35

▷

yksinkertaiset sanat											yhdyssanat			
taivutustyyppi	mallisana	leksikaalinen taajuus	log(leksikaalinen taajuus)	mediaani(lemmataajuus)	log(mediaani(lemmataajuus))	max(lemmataajuus)	log(max(lemmataajuus))	tavujen keskiarvo	hapaksien määrä	log(hapaksien määrä)	leksikaalinen taajuus	log(leksikaalinen taajuus)	mediaani(lemmataajuus)	log(mediaani(lemmataajuus))
36	(sisin)	10	2,3	2393,5	7,78	6811	8,83	2	0	0	1	0	0	0
37	(vasen)	1	0	6805	8,83	6805	8,83	2	0	0	0	0	0	0
38	(nainen)	3287	8,1	44	3,78	215711	12,28	4,2	161	5,08	5643	8,64	11	2,4
39	(vastaus)	2622	7,87	22	3,09	143495	11,87	2,9	175	5,16	4333	8,37	16	2,77
40	(kalleus)	2169	7,68	21	3,04	69915	11,16	3,81	119	4,78	2046	7,62	13	2,56
41	(vieras)	440	6,09	129	4,86	40727	10,61	2,85	11	2,4	810	6,7	14	2,64
42	(mies)	1	0	169514	12,04	169514	12,04	1	0	0	280	5,63	97	4,57
43	(ohut)	12	2,48	242,5	5,49	21759	9,99	2	0	0	24	3,18	16,5	2,8
44	(kevät)	1	0	44953	10,71	44953	10,71	2	0	0	4	1,39	120	4,79
45	(kahdeksas)	15	2,71	3091	8,04	50288	10,83	2,4	0	0	0	0	0	0
46	(tuhat)	1	0	33385	10,42	33385	10,42	2	0	0	1	0	10823	9,29
47	(kuollut)	31	3,43	514	6,24	15387	9,64	3,7	0	0	30	3,4	1,5	0,41
48	(hame)	828	6,72	88	4,48	131977	11,79	2,56	39	3,66	3854	8,26	16	2,77
49	(askel)	22	3,09	87	4,47	5752	8,66	2	1	0	76	4,33	3	1,1

Liite 2. Produktiivisuusasteet (P kategoriaehdollinen, P* hapaksiehdollinen). Sarakkeet ovat laskevassa järjestyksessä (kärjessä suurin produktiivisuusaste). (E-merkintä tarkoittaa, että luku on kirjoitettu eksponentiaalisessa muodossa: E:tä seuraava luku ilmaisee, monennella kymmenen potenssilla perusluku kerrotaan. Esimerkiksi 7,67E-05 = 7,67 * 10^{-5} = 0,0000767.)

taivutustyyppi	**mallisana**	**P**	**taivutustyyppi**	**mallisana**	**P***	**taivutustyyppi**	**mallisana**	**P (estimoitu)**	**taivutustyyppi**	**mallisana**	**P* (estimoitu)**
20	(filee)	0,000436427	5	(risti)	0,195099	49	(askel)	0,1125706	5	(risti)	0,2432327
33	(kytkin)	0,000185114	39	(vastaus)	0,1649387	20	(filee)	0,07815448	39	(vastaus)	0,240896
11	(omena)	0,000150921	38	(nainen)	0,1517436	8	(nalle)	0,07214964	38	(nainen)	0,2064508
4	(laatikko)	0,000129679	40	(kalleus)	0,1121583	39	(vastaus)	0,0647415	40	(kalleus)	0,1587308
40	(kalleus)	0,000118412	10	(koira)	0,06409048	3	(valtio)	0,05439293	10	(koira)	0,0740357
14	(solakka)	9,10E-05	12	(kulkija)	0,03675778	40	(kalleus)	0,05156883	12	(kulkija)	0,05004873
49	(askel)	7,67E-05	48	(hame)	0,03675778	33	(kytkin)	0,04987928	2	(palvelu)	0,04770159
34	(onneton)	7,38E-05	1	(valo)	0,03581527	4	(laatikko)	0,04718551	48	(hame)	0,04652883
39	(vastaus)	6,53E-05	2	(palvelu)	0,03204524	28	(kynsi)	0,04498014	6	(paperi)	0,040107
5	(risti)	5,80E-05	6	(paperi)	0,03204524	38	(nainen)	0,04425911	3	(valtio)	0,03604737
3	(valtio)	4,61E-05	3	(valtio)	0,0273327	48	(hame)	0,03959839	1	(valo)	0,0334773
6	(paperi)	4,03E-05	9	(kala)	0,02450518	5	(risti)	0,03858595	9	(kala)	0,03248373

taivutustyyppi	mallisana	P	taivutustyyppi	mallisana	P*	taivutustyyppi	mallisana	P (estimoitu)	taivutustyyppi	mallisana	P* (estimoitu)
38	(nainen)	3,94E-05	33	(kytkin)	0,02073516	17	(vapaa)	0,03636841	33	(kytkin)	0,02357103
12	(kulkija)	3,56E-05	34	(onneton)	0,01508011	12	(kulkija)	0,03265541	4	(laatikko)	0,01761078
2	(palvelu)	2,98E-05	4	(laatikko)	0,0131951	14	(solakka)	0,03034773	41	(vieras)	0,01123159
48	(hame)	2,26E-05	41	(vieras)	0,01036758	11	(omena)	0,0294225	14	(solakka)	0,009948382
17	(vapaa)	1,81E-05	14	(solakka)	0,008482564	2	(palvelu)	0,02868078	34	(onneton)	0,009607029
41	(vieras)	1,52E-05	8	(nalle)	0,003770028	10	(koira)	0,02493824	8	(nalle)	0,005119392
13	(katiska)	1,41E-05	13	(katiska)	0,002827521	6	(paperi)	0,02474799	49	(askel)	0,005092951
10	(koira)	1,37E-05	20	(filee)	0,002827521	36	(sisin)	0,02216633	20	(filee)	0,004880011
8	(nalle)	1,31E-05	15	(korkea)	0,001885014	9	(kala)	0,02100028	13	(katiska)	0,003681564
1	(valo)	8,23E-06	17	(vapaa)	0,001885014	45	(kahdeksas)	0,01807037	11	(omena)	0,001920665
9	(kala)	5,00E-06	11	(omena)	0,000942507	41	(vieras)	0,01798764	17	(vapaa)	0,00190959
15	(korkea)	3,66E-06	18	(maa)	0,000942507	13	(katiska)	0,01695613	15	(korkea)	0,001181579
28	(kynsi)	2,75E-06	28	(kynsi)	0,000942507	22	(parfait)	0,01495115	28	(kynsi)	0,000893641
31	(kaksi)	2,53E-06	31	(kaksi)	0,000942507	1	(valo)	0,01415144	32	(sisar)	0,000808889
18	(maa)	1,01E-06	49	(askel)	0,000942507	34	(onneton)	0,01251347	47	(kuollut)	0,000442268
7	(ovi)		7	(ovi)		47	(kuollut)	0,01005333	45	(kahdeksas)	0,000384656
16	(vanhempi)		16	(vanhempi)		43	(ohut)	0,01002159	36	(sisin)	0,000314563
19	(suo)		19	(suo)		32	(sisar)	0,008028169	22	(parfait)	0,00023339
21	(rosé)		21	(rosé)		15	(korkea)	0,004985769	43	(ohut)	0,00017066
22	(parfait)		22	(parfait)		16	(vanhempi)	0,001668556	16	(vanhempi)	4,50E-05
23	(moni)		23	(moni)		21	(rosé)	0,000783582	7	(ovi)	4,26E-05
24	(uni)		24	(uni)		7	(ovi)	0,000267857	21	(rosé)	1,56E-05
25	(toimi)		25	(toimi)		18	(maa)	2,81E-06	18	(maa)	2,00E-07
26	(pieni)		26	(pieni)		27	(käsi)	2,53E-08	27	(käsi)	7,19E-10
27	(käsi)		27	(käsi)		25	(toimi)	1,42E-08	25	(toimi)	1,61E-10
29	(lapsi)		29	(lapsi)		26	(pieni)	6,59E-10	24	(uni)	9,74E-11
30	(veitsi)		30	(veitsi)		24	(uni)	9,74E-11	26	(pieni)	2,71E-11
32	(sisar)		32	(sisar)		23	(moni)	7,66E-12	23	(moni)	9,78E-14
35	(lämmin)		35	(lämmin)		19	(suo)		19	(suo)	
36	(sisin)		36	(sisin)		29	(lapsi)		29	(lapsi)	
37	(vasen)		37	(vasen)		30	(veitsi)		30	(veitsi)	
42	(mies)		42	(mies)		31	(kaksi)		31	(kaksi)	
43	(ohut)		43	(ohut)		35	(lämmin)		35	(lämmin)	
44	(kevät)		44	(kevät)		37	(vasen)		37	(vasen)	
45	(kahdeksas)		45	(kahdeksas)		42	(mies)		42	(mies)	
46	(tuhat)		46	(tuhat)		44	(kevät)		44	(kevät)	
47	(kuollut)		47	(kuollut)		46	(tuhat)		46	(tuhat)	

▷

INDICES OF PRODUCTIVITY IN FINNISH INFLECTION

The aim of the present study is to describe how productivity is manifested in the inflection of Finnish nominals (i.e. nouns and inflectionally noun-like syntactic categories), when productivity is analysed with the help of different linguistic variables. Our goals also include the quantification of productivity in morphological systems that are typologically similar to that of Finnish. In other words, we attempt to obtain responses to such questions as how the productivity of an inflectional category is reflected in the Finnish inflectional system, and how many indices of productivity there are. In addition, we aim to analyse potential, quantifiable interaction between these various indices.

In each inflectional category, we examine the relationship between category size and frequency of use (lemma frequency) by using regression analysis. The variables include lexical factors (e.g. median of lemma frequency, number of hapaxes) as well as a phonological factor, namely the number of syllables. The reliability of the present results is enhanced by the relatively large size of the data: we employ the paradigm classification found in two extensive monolingual dictionaries, viz. *Suomen kielen perussanakirja* and *CD-Perussanakirja* with their 49 nominal paradigms, circa 25,000 lexemes and 52,000 compounds, as well as the corpora of the Language Bank (of Finland; www.csc.fi) containing over 130 million running words.

Our main finding is that there exists a strong correlation between paradigm size and the frequency of use of a given word in Finnish (and, presumably, in all typologically similar languages too) as follows: productive paradigms are extensive and they are characterised by low frequency of use, while unproductive paradigms are narrow but their frequency of use is high. In addition, we applied the formulae developed by Harald Baayen (e.g. Baayen 2003) to the frozen paradigms and to those that are not unambiguously productive. ■

Kirjoittajien yhteystiedot (address):
Joensuun yliopisto
Yleinen kielitiede
PL 111
80101 Joensuu
Sähköpostit: *alexandre.nikolaev@joensuu.fi*
jussi.niemi@joensuu.fi