Tampereen yliopisto

Veera Nurminen

EDUSKUNNAN TÄYSISTUNTOJEN PUHEENVUOROJEN KIELENKÄYTÖN MUUTOKSET 1980-LUVULTA NYKYPÄIVÄÄN

Kandidaatintyö
Informaatioteknologian ja viestinnän tiedekunta
Tarkastaja: Prof. Okko Räsänen
Toukokuu 2025

TIIVISTELMÄ

Veera Nurminen: Eduskunnan täysistuntojen puheenvuorojen kielenkäytön muutokset 1980-luvulta nykypäivään
Kandidaatintyö
Tampereen yliopisto
Tieto- ja sähkötekniikan kandidaatinohjelma, sähkötekniikka
Toukokuu 2025

Tämän kandidaatintyön tavoitteena oli selvittää, miten suomen kielen käyttö eduskunnan täysistunnoissa on muuttunut vuosien 1980–2023 välisenä aikana. Aineistona käytettiin laajaa eduskunnan täysistuntojen transkriptioaineistoa, josta analyysiin otettiin mukaan yhteensä noin 150 000 suomenkielistä puheenvuoroa 20 eri vuodelta.

Puheenvuoroista analysoitiin valikoituja lingvistisiä ominaisuuksia hyödyntäen luonnollisen kielen käsittelyn menetelmiä ja siihen soveltuvia valmiita työkaluja. Ominaisuuksien pohjalta laskettiin kielen rakennetta kuvaavia tunnus- ja suhdelukuja, joita verrattiin vuosilukuihin korrelaatioanalyysin avulla. Näin voitiin tunnistaa kielen ominaisuuksissa tapahtuneita ajallisia muutoksia sekä piirteitä, jotka ovat pysyneet ennallaan.

Analyysi keskittyi erityisesti eduskunnan puheenvuorojen leksikaalisiin valintoihin, sanaluokkien suhteisiin puheenvuoroissa sekä lauserakenteiden kompleksisuuteen. Tutkimuksen tulosten perusteella ajallisia muutoksia on tapahtunut erityisesti puheissa käytetyssä sanastossa. Keskimääräinen sanapituus on hieman lyhentynyt ja kielen variaatio on ajan myötä vähentynyt. Substantiivien ja verbien osuudet puheenvuoroissa ovat keskimäärin vähentyneet vuosien varrella, kun taas pronominien ja konjunktioiden osuudet ovat kasvaneet. Lauserakenteiden kompleksisuudessa tai puheenvuorojen keskimääräisissä pituuksissa ei ole tulosten perusteella tapahtunut huomattavaa muutosta tarkastelujanjakson aikana.

Tämän työn analyysimenetelmiä voidaan soveltaa myös laajempaan kielen muutosten tutkimukseen tai tarkempien tarkastelutasojen tutkimukseen, kuten yksittäisten puolueiden kielen tarkasteluun. Työssä käytetyt menetelmät mahdollistavat kielen muutosten tarkastelun automaattisesti suurista tekstikorpuksista.

Avainsanat: luonnollisen kielen käsittely, eduskuntatutkimus, automaattinen tekstianalyysi, poliittinen kieli, kielen muutos

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -ohjelmalla.

ALKUSANAT

Tämä työ osoitti minulle olevani oikealla polulla. Pääsin hyödyntämään sähkötekniikan – tarkemmin signaalinkäsittelyn ja koneoppimisen – opiskeluissa mukaani tarttuneita oppeja. Lisäksi pääsin oppimaan paljon uutta erityisesti luonnollisen kielen käsittelystä, eduskuntakäytännöistä sekä tietenkin tutkimuksen tekemisestä. Tähän pisteeseen pääseminen on vaatinut vuosien työn, enkä olisi tänne asti päätynyt ilman ohjausta ja tukea.

Haluan kiittää professori Okko Räsästä sekä tutkija Kirsi Sandbergia mahtavasta kandidaatintyön aiheesta sekä kallisarvoisesta tuesta, ohjeista ja palautteesta koko prosessin aikana. Vaikka aihepiirissä oli minulle paljon myös uutta, teidän neuvojenne ja asiantuntevan apunne ansiosta sain hyvän mahdollisuuden tutustua uusiin aihealueisiin samalla kun pääsin hyödyntämään aikaisempaa oppimaani.

Lisäksi, suuri kiitos perheelleni ja ystävilleni tuestanne ja kannustuksestanne. Kiitos myös mielenkiinnostanne tätä työtä kohtaan, joka loi minulle (positiivista) painetta saada kandidaatintyö ja tutkimus nopeasti etenemään. Tämän työn pohjalta on hyvä jatkaa diplomiinsinöörin tutkintoa kohti.

Tampereella, 15. toukokuuta 2025

Veera Nurminen

TEKOÄLYN KÄYTTÖ TÄSSÄ TYÖSSÄ

Opinnäytteessäni on käytetty tekoälysovelluksia:

- ☐ Ei
- ☒ Kyllä

Ilmoitukseni mukaan olen käyttänyt opinnäytteessäni tutkielmaprosessin aikana seuraavia tekoälysovelluksia: ChatGPT

Tekoälysovellusten nimet ja versiot: ChatGPT versio GPT-4o

Tekoälyn käyttötarkoitus

Tekoälyä on hyödynnetty kandidaatintyön oikeinkirjoituksen tarkistamiseen syöttämällä sille tekstiosioita ja pyytämällä analyysi oikeinkirjoituksesta.

Osiot, joissa tekoälyä on käytetty

Tekoälyä on hyödynnetty oikeinkirjoituksen tarkistamiseen kaikissa tekstikappaleissa.

Riskien tiedostaminen

Olen tietoinen siitä, että olen täysin vastuussa koko opinnäytteeni sisällöstä, mukaan lukien osat, joiden tuottamisessa on hyödynnetty tekoälyä, ja hyväksyn vastuun mahdollisista tästä seuranneista eettisten ohjeiden rikkomuksista.

SISÄLLYSLUETTELO

1. Johdanto 1
2. Taustaa 3
2.1 Luonnollisen kielen käsittelyn menetelmistä yleisesti 4
2.1.1 Tokenisointi 5
2.1.2 Lemmatisointi 6
2.1.3 Riippuvuusanalyysi ja sanaluokan merkkaus 6
2.2 Kielentunnistus 8
2.3 Pearson-korrelaatio 9
3. Kielenkäytön ajallinen tutkimus ja menetelmät 10
3.1 Aineiston esittely 10
3.2 Aineiston esikäsittely 11
3.2.1 Kielensuodatus 12
3.2.2 Kielenkäsittely 13
3.3 Puheen ominaisuuksien mittaaminen 14
4. Tulokset ja niiden tarkastelu 19
4.1 Tulokset 19
4.2 Pohdintaa 24
4.3 Tutkimuksen rajoitteet 25
4.4 Eduskunnan kielenkäytön tutkimus jatkossa 27
5. Yhteenveto 28
Lähteet 30
Liite A: Esimerkki eduskunnan täysistunnon puheenvuoron transkriptiosta 33

KUVALUETTELO

2.1 Esimerkki riippuvuusanalyysin tuloksista virkkeelle ”Eduskunnan seuraava täysistunto on huomenna tiistaina kello 10.” 7

3.1 Havainnollistus puheenvuorojen käsittelyvaiheista 11
3.2 Histogrammi, joka esittää muun kuin suomenkielisten puheenvuorojen osuudet kaikista puheenvuoroista. 13

4.1 Puheenvuorojen pituuksien ja sanojen keskipituuksien ajallisia muutoksia havainnollistavat kuvaajat. 20
4.2 Juuresta riippuvien sanojen määrän kehitykseen liittyvät kuvaajat. 21
4.3 Eri sanaluokkien osuuksia kuvaavat tuloskuvaajat. 22
4.4 Vasemmalla on sanaston monimuotoisuuden mittarin TTR-luvun (type/token ratio) analyysin tuloskuvaaja, ja oikealla sanan ja sen lemmatisoidun vastineen pituussuhteen tutkimuksen tuloskuvaaja. 23
4.5 Kuvaajat yhdyssanojen osuuksien analyysin tueksi. 24

TAULUKKOLUETTELO

2.1 Universaalit sanaluokkatunnukset (universal POS tags) (de Marneffe et al. 2021) 8

3.1 Esimerkki Stanza-analyysin tuloksista virkkeelle ”Eduskunnan seuraava täysistunto on huomenna tiistaina kello 10.” 14
3.2 Kielen ominaisuudet, joiden ajallisia muutoksia eduskunnan puheenvuoroissa analysoitiin tässä tutkimuksessa. 15
3.3 Kaikki tutkimuksessa tutkitut sanaluokat. 17

1. JOHDANTO

Parlamenteissa pidetyt puheenvuorot ovat ikkuna kunkin aikauden poliittisiin ja kielellisiin normeihin sekä käytänteisiin. Parlamenttien puheenvuoroista ja niiden kielestä heijastuvat yhteiskunnan senhetkinen tila ja arvot. Puheiden sisältö ja kieli sekä niiden ajallinen muutos ovat siten kiinnostavia tutkimuskohteita.

Eduskunnan täysistuntojen puheenvuorot tarjoavat kattavan aineiston poliittisen puheen ja yleisestikin kielenkäytön tutkimukseen, sillä aineistoa on saatavilla yli sadan vuoden ajalta. Eduskunta-aineisto onkin yksi laajimmista julkisista suomenkielisistä tekstikorpuksista, mikä tekee siitä arvokkaan lähteen suomen kielen muutosten pitkän aikavälin tarkasteluun. Eduskunnan puheenvuorojen tutkimus voi kertoa ajankohtaisen poliittisen kielenkäytön lisäksi ajan yleiskielestä. Vastaavaa suomen kielenkäytön tutkimusta eduskunta-aineistosta ei ole aiemmin laajalti tehty, joskin aineistona se on suosittu tutkimuskohde. Aikaisempi tutkimus on keskittynyt enemmän kielen semantiikkaan kuin itse kielenkäyttöön ja kielellisiin ominaisuuksiin.

Teknologian kehitys on avannut uusia mahdollisuuksia automaattiseen kielen käsittelyyn suurten aineistojen pohjalta. Luonnollisen kielen käsittelyn menetelmät ja niiden tekniset toteutukset mahdollistavat muun muassa lauserakenteiden, sanojen perusmuotojen ja sanaluokkien automaattisen ja koneellisen määrittämisen. Näiden menetelmien avulla kielen piirteitä voidaan analysoida valtavista tekstimassoista, joiden manuaalinen analyysi olisi äärimmäisen epäkäytännöllistä.

Tässä tutkimuksessa tarkastellaan, miten suomen kielen käyttö on muuttunut Suomen eduskunnan täysistunnoissa vuosien 1980–2023 välillä. Käytännössä puheista poimitaan ja analysoidaan valikoituja kielen ominaisuuksia, joiden pohjalta aineistosta lasketaan kielen rakennetta kuvaavia tunnus- ja suhdelukuja. Laskettuja tunnuslukuja verrataan vuosilukuihin korrelaatioanalyysin keinoin, jotta voidaan tunnistaa kielen ominaisuuksien ajallisia muutoksia tai vastaavasti pysyvyyttä. Valitut ominaisuudet painottuvat muun muassa puheiden leksikaalisiin valintoihin, sanaluokkien suhteisiin ja lauserakenteiden kompleksisuuteen. Vaikka tässä työssä keskitytään eduskunnan kieleen yleisemmällä tasolla, voi menetelmiä soveltaa myös tarkempiin analyyseihin, kuten yksittäisten puolueiden tai kansanedustajien kielen analysointiin.

Luvussa 2 pohjustetaan tutkimuksen taustatietoja sekä sen kannalta olennaisia käsitteitä

yleisesti. Lisäksi perehdytään työssä hyödynnettävien luonnollisen kielen käsittelyn menetelmien teoriaan. Aineisto ja sen esikäsittelymenetelmät esitellään luvussa 3. Kyseisessä luvussa käydään myös läpi analyysiin valitut kielen ominaisuudet ja ominaisuuksien mittaamiseen liittyvät käytännön asiat. Tutkimuksen tulokset ja niiden tarkempi analyysi esitetään luvussa 4. Lopuksi luvussa 5 esitellään vielä tutkimuksen johtopäätökset ja yhteenveto.

2. TAUSTAA

Eduskunnan täysistunnoissa pidetyillä suullisilla puheenvuoroilla on tiettyjä piirteitä, jotka erottavat ne arkisesta puheesta. Ne eroavat muista puheenlajeista muun muassa virallisuutensa, tavoitteellisuutensa ja yleisösuhteensa vuoksi. Puheenvuoroissa yhdistyvät keskustelu, poliittinen vaikuttaminen ja yhteiskunnallinen tiedottaminen. Puheilla on lähtökohtaisesti jokin poliittinen tarkoitusperä, joka pohjautuu paitsi puhujan omaan ideologiaan myös usein tämän edustaman puolueen linjauksiin.

Eduskunnan puheenvuorojen potentiaalinen yleisö on moninainen ja koostuu muun muassa eduskunnan muista jäsenistä, mediasta, eri alojen asiantuntijoista sekä eritaustaisista kansalaisista (Voutilainen 2025). Puheenvuorot voivat olla spontaania keskustelua, valmiin puheen ääneen lukemista tai jotakin siltä väliltä. Kieli on muodollisempaa kuin arkipuhe ja siinä korostuvat argumentointi, debatti sekä vakiintuneet roolikohtaiset käytännöt. Esimerkiksi kansanedustajan puheenvuoro alkaa usein sanoilla "Arvoisa puhemies". Myös puhemiehen puheenvuoroissa toistuvat tietyt roolille ominaiset fraasit ja niiden variaatiot, joiden tarkoituksena on ohjata täysistunnon kulkua. Puhemies ei kuitenkaan osallistu varsinaiseen keskusteluun (Suomen perustuslaki 731/1999, 42 §).

Puheenvuorotyyppejä on useita, varsinaisten puheenvuorojen lisäksi esimerkiksi esittelypuheenvuoroja, ryhmäpuheenvuoroja sekä vastauspuheenvuoroja (Eduskunnan työjärjestys 40/2000, 50–52 §). Eri puheenvuorotyypeille on määritetty erilaisia kestorajoituksia, joskin Suomen perustuslain (731/1999, 31 §) mukaan kansanedustajalla on eduskunnassa oikeus puhua vapaasti kaikista keskusteltavana olevista asioista sekä niiden käsittelystä.

Ajan myötä eduskuntapuheiden kielelliset ja sisällölliset piirteet voivat muuttua esimerkiksi yhteiskunnallisten tapahtumien, poliittisen ilmapiirin tai kielen yleisen kehittymisen myötä. Muutokset voivat käytännössä olla muun muassa sanavalintojen ja lauserakenteiden ajallisia muutoksia. Näiden muutosten tutkiminen tarjoaa arvokasta tietoa poliittisen viestinnän kehityksestä ja kielen sopeutumisesta yhteiskunnallisiin muutoksiin. Eduskuntapuheiden kielellisten ominaisuuksien automatisoitu tarkastelu edellyttää kuitenkin yhtenäisiä aineistoja koneellisesti analysoitavassa muodossa.

Viime vuosina kiinnostus parlamenttien puheenvuorojen tutkimukseen on ollut suurta kansainvälisesti. Tähän on vaikuttanut muun muassa se, että parlamenttiaineistot ovat

nyt helpommin käsiteltävässä muodossa niiden digitalisoinnin myötä. Monilla valtioilla on omia kansallisia parlamenttiaineistojaan, jotka voivat sisältää esimerkiksi transkriptioita parlamenteissa pidetyistä puheenvuoroista. Ne eivät kuitenkaan aikaisemmin ole välttämättä olleet helposti koneellisesti analysoitavassa muodossa.

Kansainvälisesti parlamenttiaineistojen digitalisointia on edesauttanut muun muassa tutkimusinfrastruktuuri CLARIN (Common Language Resources and Technology Infrastructure) (*CLARIN in a Nutshell* n.d.). Lyhyesti kiteytettynä CLARIN tukee kieliaineistojen ja kieliteknologisten työkalujen saatavuutta kielentutkijoille (*CLARIN in a Nutshell* n.d.). CLARINin piirissä käynnistettiin 2010-luvun loppupuolella hanke parlamenttiaineistojen yhtenäistämiseksi (Erjavec et al. 2023). Tämän työn tuloksena syntyivät Parla-CLARIN-suositukset, joiden tavoitteena on tukea eri valtioiden parlamenttiaineistojen keskinäistä vertailua ja uudelleenkäyttöä yhtenäistämällä aineistojen käytäntöjä (Erjavec ja Pančur 2019). Parla-CLARIN-suositusten myötä käynnistettiin ParlaMint-projekti, jonka lomassa kehitetyt parlamenttiaineistot sisältävät transkriptioita 17 eri Euroopan valtion parlamenteista yhtenäisessä ja helposti käsiteltävässä muodossa (Erjavec et al. 2023).

Suomessa eduskunnan puheenvuorodataa on yhtenäistetty osana Semanttinen parlamentti -hanketta, jossa on luotu tässä työssä ainestona käytetty Parlamenttisampo (Hyvönen et al. 2024). Parlamenttisampo on alkuvuonna 2023 julkistettu eduskunnan tietokannoista ja niihin liittyvistä muista aineistoista koottu datapalvelu, jossa eduskunnan aineistot on muutettu yhtenäiseen muotoon niihin pohjautuvan tutkimuksen helpottamiseksi (Hyvönen et al. 2024). Myös Parlamenttisammon aineistot ovat saatavissa Parla-CLARIN ja ParlaMint -formaateissa, mikä mahdollistaa niiden samanaikaisen käsittelyn ja vertailun muiden maiden vastaavien aineistojen kanssa.

Parlamenttien puheenvuorojen tutkimus onkin hyvin ajankohtaista niin Suomessa kuin kansainvälisesti, sillä vasta viime vuosina on saatu käyttöön helposti koneellisesti analysoitavia aineistoja. Näiden aineistojen ansiosta parlamenttipuheenvuorojen piirteet ovat nyt tarkasteltavissa laajamittaisesti automaattisen analyysin keinoin. Parlamenttisampo ja sen kautta tutkimuksessa käytetty aineisto esitellään yksityiskohtaisemmin luvussa 3.

2.1 Luonnollisen kielen käsittelyn menetelmistä yleisesti

Kirjoitetun kielen automaattiseen analyysiin on kehitetty useita luonnollisen kielen käsittelyn (natural language processing, NLP) menetelmiä. Tässä tutkimuksessa keskeisessä roolissa on Stanza, joka on luonnollisen kielen käsittelyn Python-analyysipaketti (Qi et al. 2020). Stanza tarjoaa monikielisen neuroverkkopohjaisen kanavan (neural pipeline), joka ottaa syötteenä tekstiä ja tuottaa sen pohjalta annotaatioita, kuten tokenisoinnin, lemmatisoinnin, sanaluokka-analyysin sekä sanojen riippuvuusanalyysin (Qi et al. 2020). Tästä syystä Stanzan toimintaa käytetään esimerkkinä alla kuvatuissa luonnollisen kielen käsittelyn osaongelmissa. Nämä menetelmät ovat tärkeässä osassa niin esikäsiteltäessä

tekstiä sen analysointia varten, kuin myös tarkasteltaessa tekstistä sen kielellisiä ominaisuuksia.

2.1.1 Tokenisointi

Yksi keskeinen NLP-menetelmä on *tokenisointi*. Se on useimmiten ensimmäinen vaihe käsiteltäessä tekstiä luonnollisen kielen käsittelyn keinoilla (tekstiaineiston mahdollisen siivouksen jälkeen). Tokenisointi on prosessi, jossa teksti jaetaan pienemmiksi yksiköiksi (tokeneiksi) (Bird et al. 2009), jotka voivat olla esimerkiksi sanoja, sanan osia, lauseita tai merkkejä. Tokenisoinnin avulla teksti muutetaan muotoon, jota on helpompi jatkokäsitellä automaattisesti tietokoneen avulla.

Tokenisointimenetelmät ovat hyvin kieliriippuvaisia. Tokenisointi on haastavampaa kielille, jotka eivät käytä välilyöntejä sanojen erottamiseksi toisistaan (Jurafsky ja Martin 2025). Tästä syystä esimerkiksi kiinan kielelle on toimivampaa käyttää merkkipohjaista tokenisointia (Jurafsky ja Martin 2025). Suomen kielessä sanat on erotettu toisistaan välilyönneillä, jolloin niiden jakaminen yksiköiksi on suoraviivaisempaa. Yksi yksinkertaisimmista tavoista toteuttaa tokenisointi onkin jakaa sanat yksiköiksi välilyönnin kohdalta (Bird et al. 2009). Se ei kuitenkaan usein palvele tekstin käsittelyn tarkoituksia muun muassa siksi, että välimerkit jäävät osaksi sanayksiköitä. Sopivan menetelmän valinta riippuu usein kielen piirteistä sekä käyttötarkoituksen tuomista vaatimuksista.

Mahdollisia tokenisointimenetelmiä on useita, mutta algoritmi voi pohjautua esimerkiksi sääntöpohjaiseen tai oppivaan lähestymistapaan. Sääntöpohjaisessa tokenisoinnissa teksti jaetaan yksiköiksi valittujen sääntöjen perusteella (Jurafsky ja Martin 2025), kuten juurikin leikkaamalla teksti yksiköiksi välilyöntien kohdalta. Oppivat tai tilastolliset menetelmät perustuvat siihen, että sopiviksi yksiköiksi jakaminen opitaan aineiston pohjalta tilastollisen mallin avulla. Yksinkertainen toteutus tähän on BPE-algoritmi (byte-pair encoding), joka oppii jakamaan tekstin sopiviksi yksiköiksi tutkimalla koulutuskorpusta (Jurafsky ja Martin 2025). Lyhyesti sanottuna sen sanasto koostuu aluksi vain yksittäisistä kirjaimista, mutta iteratiivisen prosessin aikana se lisää sanastoon kirjainyhdistelmiä kirjaimista, joiden se huomaa toistuvan usein vierekkäin (Jurafsky ja Martin 2025). Stanzan tokenisointi-työkalu perustuu kielikohtaisesti Universal Dependencies v2.5 Treebank -korpuksilla koulutettuun oppivaan tokenisointimalliin (Qi et al. 2020).

Tokenisointiin voidaan hyödyntää kielen käsittelyn analyysipakettien valmiita toteutuksia. Stanzan lisäksi esimerkiksi paljon käytetyt NLTK- (Natural Language Toolkit) (Bird et al. 2009) sekä SpaCy-työkalut (Honnibal et al. 2020) sisältävät omia tokenisointi-työkalujaan. Stanzaa käytettäessä sen tokenisointialgoritmi jakaa tekstin sanat ja välimerkit omiksi yksiköikseen ja ryhmittelee yksiköt virkkeiksi (Qi et al. 2020), siten säilyttäen virkkeiden sisäiset suhteet.

2.1.2 Lemmatisointi

Lemmatisointi tarkoittaa käytännössä sanojen muuttamista niiden perusmuotoihin, jolloin sanoista pyritään poistamaan mahdolliset taivutuspäätteet (esimerkiksi *sanoi* $\rightarrow$ *sanoa*). Sanan perusmuoto tunnetaan nimellä lemma. Lemmatisointi on keskeinen osa tekstin esikäsittelyä tutkittaessa tekstiä NLP-keinoin ja sitä voidaan tarvita esimerkiksi tiettyjen kielellisten ominaisuuksien tarkastelua varten. Kehittyneimmät lemmatisointitavat vaativat sanojen kokonaisvaltaista morfologista analyysiä lemman ja taivutuspäätteiden erottamiseksi toisistaan (Jurafsky ja Martin 2025). Näiden menetelmien algoritmit voivat olla hyvinkin monimutkaisia.

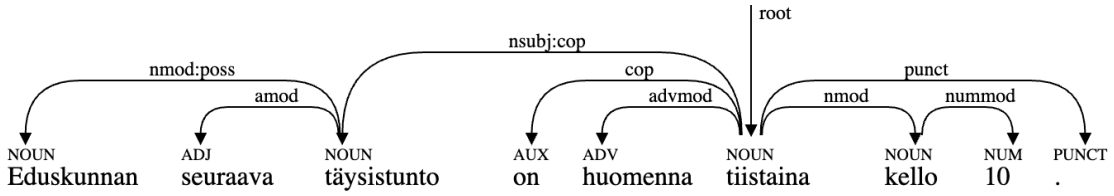
Myös lemmatisointitoteutukset voivat perustua sääntöpohjaisiin, tilastollisiin tai oppiviin menetelmiin (Akhmetov et al. 2020). Sääntöpohjaisen lemmatisoinnin haasteena on säätää olemassa olevat säännöt toimimaan uusien sanojen luokittelussa (Akhmetov et al. 2020). Lemmatisointi pelkästään ennalta määritetyn sanakirjan pohjalta voi olla epäkäytännöllistä kielen suuren variaation vuoksi (Akhmetov et al. 2020). Stanzan lemmatisointityökalu hyödyntääkin yhdistelmää sanakirjaan sekä sequence to sequence -neuroverkkoon pohjautuvaa oppivaa lemmatisointia (Qi et al. 2020).

2.1.3 Riippuvuusanalyysi ja sanaluokan merkkaus

Riippuvuusanalyysi (dependency parsing) on luonnollisen kielen käsittelyn menetelmä, jonka avulla selvitetään sanojen kieliopilliset suhteet toisiinsa virkkeen sisällä. Näiden suhteiden osapuolet ovat pääsana (head) ja riippuvainen sana (dependent) (Jurafsky ja Martin 2025). Riippuvuuspohjaisessa analyysissä pääsanat liitetään suoraan niihin sanoihin, jotka ovat niistä välittömästi riippuvaisia (Jurafsky ja Martin 2025). Riippuvuusanalyysille on olemassa useita erilaisia esitystapoja, esimerkiksi Universal Dependencies, Stanford Dependencies sekä Link Grammar (Jurafsky ja Martin 2025), jotka eroavat toisistaan muun muassa riippuvuussuhteiden määrittelytavoissa.

Tässä työssä hyödynnetään Universal Dependencies (UD) -kehyksen mukaista riippuvuusanalyysiä. UD on kehys ihmiskielten morfosyntaktiseen annotointiin, joka tarjoaa yhtenäisen tavan kuvata sanojen kieliopillisia suhteita ja piirteitä eri kielissä (de Marneffe et al. 2021). UD-kehykseen perustuvan riippuvuusanalyysin ajatuksena on, että jokaisella virkkeen sanalla on pääsana (head), josta se on suoraan riippuvainen. Virkkeellä on yksi juuri (root), josta kaikki muut sanat ovat joko suoraan tai epäsuorasti riippuvia. UD-kehyksen mukainen riippuvuuspuu (dependency tree), on kolme ehtoa täyttävä suunnattu graafi:

1. on olemassa vain yksi juurisolmu, joka ei ole riippuvainen muista solmuista,
2. muut solmut ovat kukin riippuvaisia vain yhdestä toisesta solmusta,



Kuva 2.1. *Esimerkki riippuvuusanalyysin tuloksista virkkeelle "Eduskunnan seuraava täysistunto on huomenna tiistaina kello 10." Kuvasta on nähtävissä sanojen väliset riippuvuussuhteet: kaaret yhdistävät sanat niiden pääsanoihin ja osoittavat suunnan pääsanasta riippuvaan sanaan. Kaariin on merkitty lyhenteet, jotka ilmaisevat sanojen syntaktisia suhteita. Sanojen päällä on näkyvissä niille luokitellut sanaluokat. Esimerkki on tehty hyödyntäen deptreepy-ohjelmakirjastoa, joka tarjoaa Python-työkaluja riippuvuuspuiden käsittelyyn (Ranta ja Masciolini 2023).*

3. koko rakenne on yhtenäinen (Jurafsky ja Martin 2025).

Juuri on suunnatun graafin ylin solmu, eikä se ole silloin suoraan riippuvainen muista solmuista.

UD-järjestelmän mukaiset riippuvuussuhteet voidaan visualisoida kuvassa 2.1 esitetyn kaltaisena graafina, jossa sanat ovat solmuja ja niiden väliset kieliopilliset suhteet kuvataan suunnattuina kaarina. Jokainen kaari osoittaa riippuvasta sanasta sen pääsanaan. Näitä riippuvuussuhteita kuvataan tunnuksilla, kuten "*root*", "*nsubj*" ja "*obj*", jotka kuvaavat sanan ja sen pääsanan kieliopillista suhdetta. Näitä eri syntaktisia suhteita kuvaavia universaaleja tunnuksia on yhteensä 37 (de Marneffe et al. 2021). Kuvasta 2.1 on nähtävissä suunnatut kaaret, jotka yhdistävät sanat niiden pääsanoihin, sekä tunnukset kaarien yhteydessä. Muun muassa Stanza hyödyntää Universal Dependencies (UD) -kehystä syntaktisten ja morfologisten analyysitietojen tunnistamiseksi (Qi et al. 2020).

Myös sanaluokkien tunnistaminen on olennaista tutkimuksen kannalta, sillä niiden suhteelliset osuudet eduskunnan puheenvuoroista ovat tässä tutkimuksessa yksi tarkastelun kohteista. *Sanaluokan merkkaus* (part-of-speech tagging, POS tagging) on keskeisessä osassa määrittäessä sanojen syntaktisia suhteita, sillä sanan sanaluokka antaa tietoa sen todennäköisistä ympäröivistä sanoista (Jurafsky ja Martin 2025). Kielen tutkimuksen kannalta voi olla tärkeää esimerkiksi erottaa onko lauseen kontekstissa sana "Suomalainen" kansallisuus (adjektiivi) vai sukunimi (substantiivi). Sanaluokan merkkaus ei siis ole aina yksiselitteistä. Sen päällimmäinen tavoite on kuitenkin selvittää oikea kontekstisidonnainen sanaluokka (Jurafsky ja Martin 2025).

Varhaisimmat sanaluokkien merkkausjärjestelmät pohjautuivat ennalta valittuihin sääntöihin, mutta nykyään yksi yleisimmistä lähestymistavoista on yhdistää sääntöpohjaista ja koneoppimiseen perustuvaa lähestymistapaa (Jurafsky ja Martin 2025). Sanaluokan merkkaukseen voidaan hyödyntää myös tilastollisia malleja, kuten Hidden Markov Model (HMM) tai Conditional Random Fields (CRFs) -menetelmiä (Jurafsky ja Martin 2025).

Sanaluokka voidaan ilmaista esimerkiksi Universal Dependencies -kehyksen mukaisilla universaaleilla sanaluokkatunnuksilla (UPOS), kuten kuvassa 2.1. Kaikki universaalit sanaluokkatunnukset on esitetty taulukossa 2.1.

Taulukko 2.1. *Universaalit sanaluokkatunnukset (universal POS tags) (de Marneffe et al. 2021)*

Lyhenne	**Sanaluokka**
ADJ	Adjektiivi
ADP	Adpositio
ADV	Adverbi
AUX	Apuverbi
CCONJ	Rinnastuskonjunktio
DET	Determinantti
INTJ	Interjektio
NOUN	Substantiivi
NUM	Numeraali
PART	Partikkeli
PRON	Pronomini
PROPN	Erisnimi
PUNCT	Välimerkki
SCONJ	Alistuskonjunktio
SYM	Symboli
VERB	Verbi
X	Muu

2.2 Kielentunnistus

Kielentunnistus (language detection) tarkoittaa tekstin tai aineiston kielen automaattista tunnistamista. Se on tyypillinen esikäsittelyvaihe käsiteltäessä tekstiä NLP-menetelmien avulla. Tässä tutkimuksessa hyödynnetään kielentunnistusmenetelmiä suodattaessa eduskunta-aineistosta pois vieraskielisiä puheenvuoroja.

Yksi yleisimpiä tapoja toteuttaa kielentunnistus on hyödyntää n-grammeja. Lyhyesti kiteytettynä n-grammit ovat *n* peräkkäisen sanan tai merkin jonoja, jotka löytyvät syötetekstistä (Rao et al. 2023). Tekstin n-grammijakauman (eli suhteellisten esiintymisfrekvenssien) pohjalta voidaan luoda tilastollinen kielenluokittelumalli (Rao et al. 2023), sillä eri kielillä on omat tyypilliset n-grammijakaumansa.

Viime vuosina koneoppimiseen perustuvat kielentunnistusmenetelmät ovat yleistyneet huomattavasti (Rao et al. 2023). Naiivi Bayesin luokittelija, neuroverkot ja tukivektorial-

goritmit (SVMs) ovat yleisiä koneoppimispohjaisia lähestymistapoja ongelmaan (Rao et al. 2023). Käytännön sovelluksiin on olemassa useita valmiita kielentunnistustyökaluja, kuten langid.py (Lui ja Baldwin 2012) ja langdetect (Danilák 2014).

2.3 Pearson-korrelaatio

Kahden muuttujan välisen tilastollisen riippuvuuden voimakkuutta voidaan tarkastella Pearsonin korrelaation avulla. Muuttujien välillä on lineaarinen yhteys, mikäli toisen muuttujan kasvaessa myös toinen muuttuja joko kasvaa tai pienenee (Kestilä-Kekkonen n.d.). Tässä tutkimuksessa Pearsonin korrelaatiota hyödynnetään tutkittaessa puheenvuorojen ajankohdan (vuosiluvun) ja puheenvuoroista mitattujen kielen ominaisuuksien välisiä yhteyksiä.

Pearsonin korrelaation avulla aineistosta lasketaan korrelaatiokerroin r, joka voi saada arvoja välillä -1 ja 1. Arvoa 1 lähestyvä korrelaatiokerroin kertoo positiivisesta yhteydestä, mikä tarkoittaa, että molempien muuttujien arvot kasvavat samassa suhteessa (Kestilä-Kekkonen n.d.). Vastaavasti arvoa -1 lähestyvä korrelaatiokerroin tarkoittaa, että toisen muuttujan arvojen kasvaessa toisen muuttujan arvot vähenevät samassa suhteessa (Kestilä-Kekkonen n.d.). Mitä lähempänä nollaa korrelaatiokerroin on, sitä heikommasta yhteydestä on kyse. Jos korrelaatiokerroin on nolla, muuttujien välillä ei ole lineaarista yhteyttä (Kestilä-Kekkonen n.d.).

Pearsonin korrelaatiokerroin voidaan määrittää Pythonin SciPy-kirjaston (Virtanen et al. 2020) pearsonr-funktiolla. Funktion avulla määritetään korrelaatiokertoimen r lisäksi p-arvo, jota voidaan hyödyntää analysoitaessa sitä kuinka todennäköistä on saada havaittu tai vahvempi korrelaatio sattumalta, jos todellinen korrelaatio muuttujien välillä olisi nolla (Berman 2016). Tyypillisesti p-arvon merkittävyyden rajaksi määritellään $0,05$. Jos p-arvo on tällöin alle $0,05$, voidaan päätellä, ettei havaittu yhteys ole todennäköisesti sattumaa. Käytännössä tämä tarkoittaa, että jos muuttujien välillä ei olisikaan todellista yhteyttä koko populaatiossa, niin havaittu korrelaatio esiintyisi sattumalta alle 5%:n todennäköisyydellä.

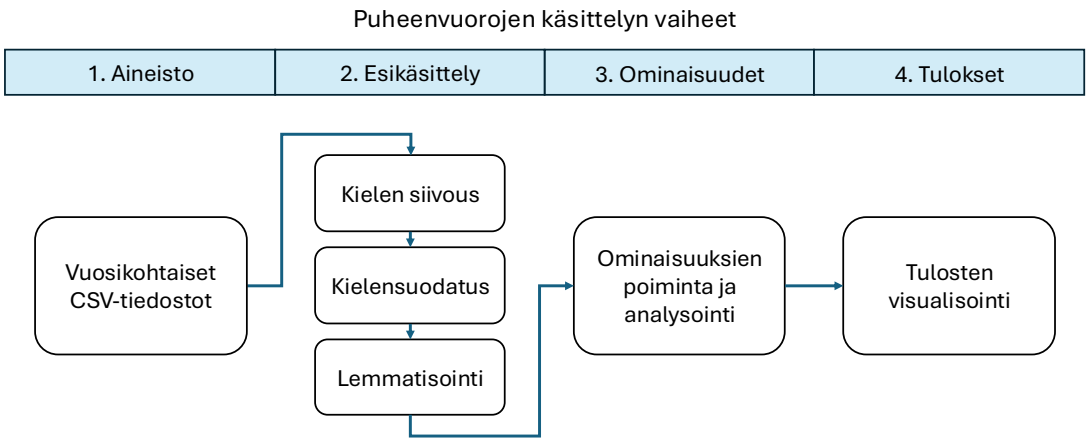
3. KIELENKÄYTÖN AJALLINEN TUTKIMUS JA MENETELMÄT

3.1 Aineiston esittely

Työssä hyödynnetään Semanttinen parlamentti -hankkeen osana luotua Parlamenttisammon tarjoamaa tekstiaineistoa (Hyvönen et al. 2021). Semanttinen parlamentti oli vuosina 2020–2022 Suomen Akatemian rahoittama hanke, jossa kehitettiin eduskunnan täysistuntojen puheenvuoroista sekä yli 2 800 eduskunnassa puhuneesta henkilöstä linkitetyn avoimen datan (Linked Open Data) infrastruktuuri eli Parlamenttisampo (Hyvönen et al. 2024). Parlamenttisampo julkaistiin alkuvuodesta 2023 ja se on osa Sampo-järjestelmien sarjaa (Hyvönen et al. 2021). Siihen on koottu lähes kaikki eduskunnan täysistuntojen puheenvuorot ja niihin liittyviä muita tietoja vuodesta 1907 lähtien aina nykypäivään saakka. Parlamenttisammon dataa ja palveluita on saatavilla muun muassa CSV-muodossa, Parla-CLARIN-formaatin mukaisessa XML-muodossa ja ParlaMint-muotoisena alikorpuksena (Hyvönen et al. 2024).

Tätä tutkimusta varten aineisto on ladattu Parlamenttisampo.fi-portaalin kautta CSV-tiedostoina. Aineistoksi valittiin eduskunnan täysistuntojen puheenvuoroja jokaisen vuosikymmenen 1980–2020 alusta neljältä ensimmäiseltä vuodelta, eli esimerkiksi vuosilta 1980–1983. Parlamenttisampo.fi-portaalin kautta haettiin vuosikohtaisesti kaikki vuoden puheenvuorot yhteen tiedostoon, mutta portaalin kautta ladatut CSV-tiedostot sisältävät maksimissaan 10 000 puheenvuoroa, jolloin sen ylittävät puheenvuororivit rajautuvat automaattisesti pois. Yksi rivi ladatusta CSV-tiedostosta sisältää aina yhden puheen sisällön sekä tietoa muun muassa puheenvuoron puhujasta, tämän puolueesta sekä puheen päivämäärästä. Yhteensä aineisto sisältää noin 200 000 puheenvuoroa ennen minkäänlaista esikäsittelyä.

Täysistuntojen puheenvuorot julkaistaan transkriptioina, eli puheenvuorot on muunnettu puheesta tekstimuotoon. Transkriptiot on luotu eduskunnan kansliassa systemaattisen ja yksityiskohtaisen kirjaamisohjeen mukaisesti, jotta puheiden käsittely olisi mahdollisimman tasa-arvoista ja systemaattista puhujasta ja transkription tekijästä riippumatta (Kirjo 2021, Voutilainen 2023 mukaan). Puhuttu puhe ja sen pohjalta kirjattu transkriptio kuitenkin eroavat toisistaan jossain määrin.



Kuva 3.1. *Havainnollistus puheenvuorojen käsittelyvaiheista*

Kirjaamisen yhteydessä tehdään fonologisia, morfologisia ja syntaktisia muutoksia, sekä muun muassa lipsahduksia ja änkytyksiä editoidaan (Voutilainen 2023). Puheenvuoroja muokataan noudattamaan yleiskielen normeja. Yleiskielellä viitataan kaikkia suomenkielisiä ihmisiä yhdistävään kielimuotoon, eli kielen paikalliset erot – kuten murteet – eroavat yleiskielestä (Hiidenmaa 2005). Muutosten ja editoinnin tarkoituksena on tukea tekstimuotoisen puheenvuoron luettavuutta ja ymmärrettävyyttä säilyttäen puheen autenttisuus mahdollisimman hyvin (Voutilainen 2023). Tarkoitus ei siis ole muokata puhetta minkäänlaiseen standardimuotoon tai ”paremmaksi” kieleksi, vaan säilyttää sen piirteitä ja variaatiota löytäen tasapaino autenttisuuden ja luettavuuden välillä (Voutilainen 2023).

Puheiden systemaattinen kirjaaminen mahdollistaa kielen muutosten tarkastelun ajan saatossa transkriptioiden pohjalta. Tämän tutkimuksen kannalta on oleellista huomioida muun muassa se, että kielen syntaktisia rakenteita editoidaan kirjaamisvaiheessa suhteellisen kevyesti ja puheen leksikaalisia valintoja ei yleisesti ottaen muokata (Voutilainen 2023). Liitteessä A on esitetty esimerkki eduskunnan täysistunnossa pidetyn puheenvuoron transkriptiosta.

3.2 Aineiston esikäsittely

Joitakin puheenvuoroista löytyviä ominaisuuksia on syytä tutkia alkuperäisestä transkriptiosta, mutta toisia varten sisältöä on tarpeen esikäsitellä. Kuvassa 3.1 on havainnollistettu puheenvuorojen käsittelyn vaiheet. Pääpiirteissään esikäsittelyvaiheessa ensin poistetaan erikoismerkit ja muutetaan teksti pieniksi kirjaimiksi sekä suodatetaan muut kuin suomenkieliset puheenvuorot pois. Sitten puheenvuorot lemmatisoidaan ja käsitellään muutenkin Stanzan avulla. Tässä osiossa avataan esikäsittelyn vaiheita yksityiskohtaisemmin.

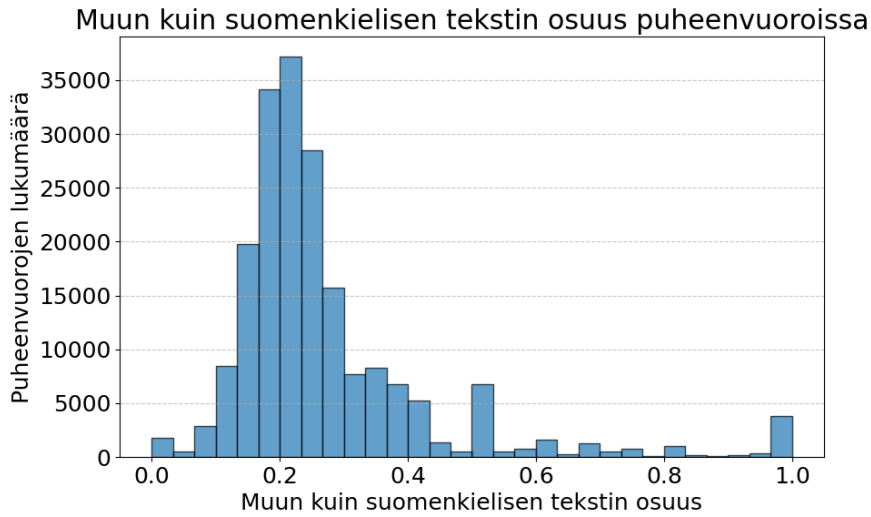
3.2.1 Kielensuodatus

Suomessa on kaksi virallista kieltä: suomi ja ruotsi. Muun kuin suomen tai ruotsin kielen osuus puheenvuoroissa on aikaisemman kielijakaumaan liittyvän tutkimuksen perusteella melko vähäistä (Jauhiainen et al. 2024). Tässä tutkimuksessa muut kuin suomenkieliset puheenvuorot pyritään poistamaan tutkimusaineistosta, jotta tutkimuksessa voidaan keskittyä suomen kielen ajallisiin muutoksiin. Käytännössä kielensuodatus toteutetaan suodattamalla lähdetiedostosta kokonaan pois ne rivit, joiden puheenvuoroissa tunnistetaan langdetect-kirjaston avulla olevan valittua raja-arvoa suurempi osuus jotakin muuta kieltä kuin suomea. Langdetect on suora Python-toteutus Java-pohjaisesta language-detection-kirjastosta ja se toimii samalla periaatteella (Danilák 2014).

Language-detection, ja sitä kautta langdetect, hyödyntää naiivia Bayesin luokittelijaa kielen ennustamiseen (Nakatani 2010). Luokittelija ennustaa tekstin kielen vertaamalla tekstistä löytyviä n-grammeja sille ennalta opetettuihin kieliprofiileihin ja päättelee mitä kieltä sille annettu teksti todennäköisimmin on (Nakatani 2010). Language-detection hyödyntää merkkipohjaista n-grammimallia (Nakatani 2010). Syötetekstille määritetään todennäköisyysjakauma, joka kuvaa language-detection-kirjaston arviota siitä, millä todennäköisyydellä teksti on kutakin sen tunnistamaa kieltä. Luokittelija päätyy luokittelemaan tekstin siihen kielikategoriaan, jonka todennäköisyys on suurin. Language-detection kirjaston kielimallit, joita langdetect myös hyödyntää, on koulutettu Wikipedia-artikkelien avulla (Nakatani 2010). Langdetect tukee yhteensä 55 eri kieltä (Danilák 2014).

Kuva 3.2 esittää muiden kuin suomenkielisten sanojen osuuden jakautumista kaikissa tutkimuksen aineistona käytetyissä puheenvuoroissa. Histogrammista nähdään, että suurin osa puheenvuoroista sisältää noin 10-30% joksikin muuksi kuin suomeksi tunnistettua kieltä. Tämä selittyy osittain sillä, että monessa puheenvuorossa esiintyy sekä ruotsin- että suomenkielistä sisältöä, mutta myös luokitteluvirheitä voi tapahtua. Histogrammin pohjalta kielensuodatuksessa käytettäväksi raja-arvoksi on valittu 30% (0,30). Raja-arvoa testattiin myös käytännössä ja se osoittautui toimivaksi kompromissiksi suomenkielisten ja muunkielisten puheenvuorojen erottelussa.

Valittu 30% raja-arvo suodattaa pois lähinnä ne puheenvuorot, joissa muunkieliseksi luokitellun sisällön osuus on erityisen korkea. Tällöin suomenkieliset puheenvuorot säilyvät suurelta osin, ja kokonaan tai suurimmaksi osaksi muunkieliset puheenvuorot rajautuvat pois. Kielensuodatuksen jälkeen tutkimusaineisto sisältää yhteensä noin 150 000 puheenvuoroa, eli noin 50 000 puheenvuoroa rajautuu pois tutkimusaineistosta kielensuodatuksen yhteydessä.



Kuva 3.2. *Histogrammi, joka esittää muun kuin suomenkielisten puheenvuorojen osuudet kaikista puheenvuoroista. X-akselilla on vieraskielisen tekstin suhteellinen osuus ja Y-akselilla kyseiseen osuuteen kuuluvien puheenvuorojen lukumäärä. Valtaosa puheenvuoroista sisältää langdetectin kieliluokitteluun perustuen 10-30% muuta kuin suomenkielistä tekstiä.*

3.2.2 Kielenkäsittely

Kielenkäsittelyvaiheessa hyödynnetään Stanzan suomen kielelle tarkoitettuja tokenisointi, lemmatisointi, sanaluokkien tunnistus ja riippuvuusanalyysiominaisuuksia. Näiden työkalujen koulutuksessa ja arvioinnissa on käytetty kielikohtaisesti Universal Dependencies v2.5 Treebank -korpuksia (Qi et al. 2020). Stanza pohjautuu Bi-LSTM-arkkitehtuuriin ja hyödyntää niin sanottua biaffiinista riippuvuusparseria, joka on syvä neuroverkkopohjainen malli syntaktisten suhteiden ja morfologisten piirteiden ennustamiseen (Qi et al. 2020). Bi-LSTM käsittelee syötetekstiä kahteen suuntaan, mikä mahdollistaa sanan vertaamisen sen kontekstiin oikeiden annotaatioiden, kuten sanaluokan määrittämiseksi (Dozat ja Manning 2017).

Taulukossa 3.1 esitetään Stanza-käsittelyn tulokset esimerkkivirkkeelle ”Eduskunnan seuraava täysistunto on huomenna tiistaina kello 10.” Jokaiselle sanalle ja välimerkille määritetään sen järjestysnumero virkkeessä (id), sen perusmuoto (lemma) ja sanaluokka (upos). Jokaiselle sanalle ja välimerkille selvitetään myös sen sanan numero, josta kyseinen sana on suoraan riippuvainen (head). Stanza määrittää tekstistä runsaasti muutakin tietoa kuin edellä mainitut, mutta tämän tutkimuksen osalta nämä ominaisuudet ovat oleellisimpia tuntea.

Ennen muita aineiston esikäsittelyvaiheita puheenvuorojen sisältöön tehdään kuitenkin yksi muutos. Kun puhuja käyttää puheessaan sanaa ”edustaja”, eduskunnan täysistuntojen puheenvuorojen transkriptioissa se on lyhennetty muotoon ”ed.” Joissakin tapauksissa virkkeiden sanariippuvuuksien tutkimiseen käytetty Stanza tunnistaa kyseisen lyhenteen

Taulukko 3.1. *Esimerkki Stanza-analyysin tuloksista virkkeelle ”Eduskunnan seuraava täysistunto on huomenna tiistaina kello 10.” Taulukon sarakkeista löytyy kunkin sanan tai välimerkin virkekohtainen järjestysnumero (id), itse sana (text), sen perusmuoto (lemma), sen universaali sanaluokka (upos) sekä sen sanan ID, josta kyseinen sana riippuu (head).*

id	text	lemma	upos	head
1	Eduskunnan	edus#kunta	NOUN	3
2	seuraava	seuraava	ADJ	3
3	täysistunto	täysistunto	NOUN	6
4	on	olla	AUX	6
5	huomenna	huomenna	ADV	6
6	tiistaina	tiistai	NOUN	0
7	kello	kello	NOUN	6
8	10	10	NUM	7
9	.	.	PUNCT	6

virkkeen lopuksi silloinkin kun virke todellisuudessa jatkuu, mikä aiheuttaa epätarkkuutta tutkimuksen tuloksiin. Tästä syystä esikäsittelyn yhteydessä lyhenne ”ed.” muutetaan sanaksi ”edustaja” kaikissa puheenvuoroissa.

Ennen kielensuodatusta kustakin puheenvuorosta poistetaan kaikki erikoismerkit ja säilytetään ainoastaan kirjaimet a-ö sekä numerot 0-9. Kaikki teksti myös muutetaan pieniksi kirjaimiksi. Tämän lisäksi esikäsittelyn yhteydessä aineistoon lisätään uutena tietona jokaisen puheenvuoron vuodet sekä puhujan roolit (puhemies, varapuhemies, muu).

Kielensuodatuksen jälkeen puheenvuorot käsitellään Stanzan avulla. Aineistoon tehdään kustakin puheenvuorosta lemmatisoidut versiot. Lemmatisointi tehdään puheenvuoroille, joista on poistettu erikoismerkit ja joiden kaikki kirjaimet on muutettu pieniksi kirjaimiksi. Stanza erottaa lemmatisoinnin yhteydessä jokaisen yhdyssanan eri osasanat toisistaan merkillä “#”. Esimerkiksi sana ”ryhmäpuheenvuorossaan” saa lemmatisointivaiheessa muodon ”ryhmä#puheen#vuoro”. Tämä on hyvin hyödyllinen ominaisuus muun muassa tutkittaessa yhdyssanojen osuuksia puheenvuoroissa. Lemmatisoidut puheenvuorot tallennetaan erikseen, jotta sekä esikäsiteltyjä alkuperäisiä että niiden lemmatisoituja versioita tekstistä voidaan analysoida erikseen yhdessä vuosiluvun ja puhujan roolin kanssa.

3.3 Puheen ominaisuuksien mittaaminen

Eduskunnan täysistunnoissa pidettyjen puheenvuorojen muutosten analysoinnissa keskityttiin useisiin rakenteellisiin ja kielellisiin ominaisuuksiin. Jokaisesta puheenvuorosta laskettiin joukko kielen rakennetta kuvaavia tunnus- tai suhdelukuja. Tunnuslukujen lukuarvoja verrattiin sitten vuosilukuihin hyödyntäen korrelaatioanalyysiä ominaisuuksien ajallisten

muutosten selvittämiseksi. Ominaisuuksien mahdolliset ajalliset muutokset tai vastaavasti muuttumattomuus kertovat poliittisen kielen kehityksestä eduskunnassa. Tutkitut ominaisuudet on esitetty taulukossa 3.2. Lähtökohtaisesti kielellisiä tunnuslukuja laskettiin puheenvuorojen tasolla ellei toisin mainita. Tässä kappaleessa mainittujen ominaisuuksien tutkimustulosten analyysi ja visuaaliset tulokset ovat eriteltynä luvussa 4.

Taulukko 3.2. *Kielen ominaisuudet, joiden ajallisia muutoksia eduskunnan puheenvuoroissa analysoitiin tässä tutkimuksessa.*

Tutkimuksessa analysoidut kielen ominaisuudet
1. Puheenvuoron pituus
2. Sanojen keskimääräinen merkkimäärä puheenvuoroa kohden
3. Virkkeen juuresta riippuvien sanojen määrä
4. Virkkeen juuresta riippuvien sanojen määrä suhteessa virkkeen sanamäärään
5. Adjektiivien, substantiivien, verbien, pronominien, interjektioiden ja konjunktioiden osuudet puheenvuoroissa
6. Sanaston monimuotoisuus type/token ration (TTR) avulla
7. Yhdyssanojen suhteellinen osuus puheenvuoroissa
8. Substantiiviyhdyssanojen osuus puheenvuorojen kaikista substantiiveista
9. Alkuperäisten ja lemmatisoitujen sanojen pituussuhde

Kielenkäytön muutoksia ajan myötä tutkittiin ensimmäisenä analysoimalla puheenvuorojen pituuksia. Puheenvuorojen pituutta tarkasteltiin laskemalla jokaisen puheenvuoron pituus merkkeinä ja tutkimalla korrelaatiota puheen esittämisvuoden ja puheenvuoron pituuden välillä. Vastaavasti laskettiin puheenvuorokohtainen keskimääräinen sanapituus merkkeinä kaikille puheenvuoroille.

Eduskunnassa käytetyn kielen kompleksisuuden kehityssuunta on tärkeä tutkimuskohde tarkasteltaessa kielen muutoksia. Kielen syntaktista kompleksisuutta tarkasteltiin tutkimalla sanariippuvuuksia sekä sanaston monimuotoisuutta puheenvuorojen virkkeissä. Käytännössä aluksi laskettiin jokaisen puheenvuoron jokaista virkettä kohden niiden sanojen määrä, jotka riippuvat suoraan virkkeen juuresta. Virkkeiden juuret ja niistä riippuvat sanat poimittiin Stanzalla toteutetuista UD-riippuvuuspuista. Tämä ominaisuus kertoo virkkeiden monimutkaisuudesta, lauserakenteiden syvyydestä ja pituudesta. Jos juuresta riippuvien sanojen määrä kasvaa, voidaan tulkita virkkeiden olevan pidempiä tai rakenteellisesti kompleksisempia. Jos taas lukema on pienempi, virkkeet voivat olla yksinkertaisempia tai lyhyempiä.

Virkkeen juuresta riippuvien sanojen määrää tutkittiin myös suhteuttamalla se virkkeen sanamäärään. Suhteellinen mittari kertoo siitä, kuinka tiiviisti juuren ympärille sanat ovat ryhmittyneet. Jokaiselle aineiston virkkeelle laskettiin suhdeluku jakamalla virkkeen juuresta riippuvien sanojen määrä virkkeen kokonaissanamäärällä. Korkeampi suhdeluku on

osoitus yksinkertaisemmasta kielestä ja sanojen tiiviimmästä ryhmittymisestä. Matalampi suhdeluku taas tarkoittaa, että virkkeissä on useampia alisteisia rakenteita ja sanoja, jotka eivät ole suoraan yhteydessä virkkeen juureen, eli kieli on kompleksisempaa. Suhteutettu mittari varmistaa sen, että mahdollinen muutos virkkeen juuresta riippuvien sanojen lukumäärässä ei johdu virkepituuden muutoksesta, vaan muutos on yhteydessä lauserakenteen monimutkaistumiseen tai yksinkertaistumiseen.

Sanaston monimuotoisuuden mittari, TTR-luku (type/token ratio), kertoo tietyn tekstikatkelman uniikkien sanojen (types) ja sen kaikkien sanojen (tokens) suhteesta (Johnson 1944). Luku lasketaan kaavan 3.1 mukaisesti jakamalla uniikkien sanojen lukumäärä tietyssä tekstiotteessa tekstin kaikkien sanojen lukumäärällä. Tutkimuksessa TTR laskettiin vuosikohtaisesti aineiston 1 000 peräkkäiselle sanalle satunnaisesti valitusta kohdasta alkaen yhteensä 1 000 kertaa. Laskenta tehtiin käyttäen puheenvuoroaineistoa, josta oli poistettu erikoismerkit ja kaikki kirjaimet muutettu pieniksi kirjaimiksi. Mitä suurempi TTR, sitä enemmän uniikkeja sanoja tekstissä on käytetty suhteessa sanamäärään. Se on merkki monimutkaisemmasta kielenkäytöstä ja suuremmasta variaatiosta käytetyssä kielessä. Pienempi TTR taas kertoo siitä, että puheessa käytetään suhteessa enemmän samoja sanoja uudelleen ja variaatiota on vähemmän.

$$\mathrm{TTR} = \frac{\text{uniikkien sanojen määrä}}{\text{kokonaissanamäärä}} \tag{3.1}$$

Sanaston mahdollisia muutoksia vuosien varrella tutkittiin myös tarkastelemalla eri sanaluokkien suhteellisia osuuksia puheenvuoroista. Kielen rakenteelliset muutokset voivat heijastua sanaluokkien osuuksiin puheessa. Siitä syystä vaihtelu sanaluokkien käytössä on hyvin tärkeä tarkastelun kohde kielenkäytön muutoksia tutkittaessa. Tässä tutkimuksessa tarkasteltiin adjektiivien, substantiivien, verbien, pronominien, interjektioiden sekä konjunktioiden osuuksia puheenvuoroissa eri aikakausina. Nämä sanaluokat valittiin sillä ne yhdessä antavat hyvän kokonaiskuvan kielen rakenteesta ja piirteistä. Analysoitavaa sanaluokkaa edustavien sanojen määrää puheenvuorossa verrattiin aina puheenvuoron kokonaissanamäärään, jotta puheiden pituuserot eivät vaikuttaisi vertailtavuuteen.

Sanaluokkien tunnistamisessa hyödynnettiin Stanzan tuottamaa universaalien sanaluokkien (UPOS) annotointia. Taulukkoon 3.3 on kerätty kaikki tutkittavat sanaluokat sekä niitä vastaavat UPOS-tunnukset, jotka huomioitiin laskettaessa sanaluokkien osuuksia. Kunkin sanaluokan osuus puheenvuorosta selvitettiin laskemalla ne sanat, joilla oli kyseinen UPOS-tunnus, ja jakamalla niiden määrä puheenvuoron kokonaissanamäärällä.

Adjektiivien osalta tutkittiin niiden sanojen osuutta, joiden sanaluokaksi Stanza tunnisti UPOS-tunnuksen ”ADJ”. Verbien osalta huomioitiin vain sanat, joiden sanaluokaksi tunnistettiin ”VERB”. Tällöin apuverbit (”AUX”) eivät olleet mukana verbien laskennassa. Substantiivien osalta otettiin tarkasteluun ainoastaan niiden sanojen osuus puheenvuo-

Taulukko 3.3. *Kaikki tutkimuksessa tutkitut sanaluokat. Kunkin sanaluokan osuus määritettiin laskemalla niiden sanojen määrä, joille Stanza tunnisti taulukon mukaisen UPOS-tunnuksen.*

Sanaluokka	**UPOS-tunnus**
Adjektiivit	ADJ
Verbit	VERB
Substantiivit	NOUN
Pronominit	PRON
Interjektiot	INTJ
Konjunktiot	CCONJ, SCONJ

roissa, joiden tunnus oli ”NOUN”. Tällöin erisnimet (”PROPN”) rajattiin tarkastelun ulkopuolelle, vaikka ne ovatkin myös substantiiveja.

Apuverbien käyttö on kytköksissä kieliopillisiin muotoihin, eivätkä ne käsitä samanlaisia sisältömerkityksiä kuin verbit yleisesti. Apuverbien määrä ei tuo varsinaista lisäarvoa verbien osuuksia tutkittaessa, joten ne rajattiin tutkimuksen ulkopuolelle. Erisnimet ovat myös hyvin selkeä omanlaisensa luokka. Ne päätettiin pitää erillään muista substantiiveista, sillä niillä viitataan lähtökohtaisesti tiettyihin henkilöihin, paikkoihin tai organisaatioihin, eikä niiden määrä kerro suoranaisesti kielen leksikaalisesta rakenteesta.

Pronominien osalta laskettiin ne sanat, joiden tunnus oli ”PRON”. Interjektioiden osalta taas ne, joiden tunnus oli ”INTJ”. Konjunktioiden osuutta laskettaessa huomioitiin sekä rinnastuskonjunktiot (”CCONJ”) että alistuskonjunktiot (”SCONJ”). Näiden molempien huomioiminen antaa kattavamman kuvan lauserakenteista ja niiden muutoksista, sillä jokainen konjunktio ilmaisee jossain määrin jotain merkityssuhdetta, esimerkiksi ehtoa, vastakohtaa, perustelua tai aikasuhdetta (*Kielitoimiston ohjepankki* n.d.[a]).

Seuraavaksi tutkittiin yhdyssanojen käytön ajallisia muutoksia. Jokaiselle puheenvuorolle laskettiin suhdeluku jakamalla sen yhdyssanojen määrä puheenvuoron kokonaissanamäärällä. Yhdyssanojen määrä määritettiin käytännössä laskemalla puheenvuorosta ne sanat, joihin Stanza merkkasi merkin ”#” tunnistamansa yhdyssanan osasanojen väliin. Koska aineistosta poistettiin esikäsittelyvaiheessa kaikki erikoismerkit ennen lemmatisointia, voidaan ”#”-merkkejä laskea luotettavasti.

Yhdyssanat ovat tavallisimmin substantiiveja, joten tutkimuksessa tarkasteltiin myös substantiiviyhdyssanojen osuutta kaikista puheenvuorojen substantiiveista. Sitä kautta voitiin selvittää ovatko mahdolliset muutokset yhdyssanojen määrässä yhteydessä substantiivien määrän muutokseen. Käytännössä puheenvuorokohtaisesti ensin laskettiin substantiivien määrä hyödyntäen Stanzan sanaluokka-annotaatioita. Substantiiveiksi laskettiin taas vain ne, joiden UPOS-tunnus oli ”NOUN”. Ei siis huomioitu erisnimiä, joiden UPOS-tunnus on ”PROPN”. Sitten puheenvuoron substantiiveista laskettiin ne, jotka oli-

vat myös yhdyssanoja. Yhdyssanojen laskemiseen hyödynnettiin tässäkin yhteydessä Stanzan merkkaamia ”#”’-merkkejä. Substantiiviyhdyssanojen osuus puheenvuoron kaikista substantiiveista saatiin jakamalla substantiiviyhdyssanat puheenvuoron kaikkien substantiivien määrällä. Tämä suhdeluku määritettiin jokaiselle aineiston puheenvuorolle.

Tutkimuksessa verrattiin puheenvuorojen alkuperäisten taivutettujen sanojen keskimääräisiä pituuksia niiden lemmatisoitujen vastineiden pituuksiin. Pituussuhde laskettiin kaavan (3.2) mukaisesti jokaiselle sanaparille ja jokaista puheenvuoroa kohden määritettiin pituussuhteiden keskiarvo. Tarkastelun kohteena oli siis kuinka paljon sanojen pituus muuttui lemmatisoinnin yhteydessä ja onko pituussuhteessa havaittavissa ajallisia trendejä.

$$\text{pituussuhde} = \frac{\text{alkuperäisen sanan merkkimäärä}}{\text{lemmatisoidun sanan merkkimäärä}} \tag{3.2}$$

Pituussuhteen ollessa lähellä lukua yksi, alkuperäisten ja vastaavien lemmatisoitujen sanojen pituus on lähes sama. Pituussuhteen ollessa enemmän kuin yksi, alkuperäiset sanat ovat merkkimäärältään pidempiä kuin lemmatisoidut sanat, mikä on odotettavaa sanojen taivutusmuotojen aiheuttaman suuremman merkkimäärän vuoksi. Jos pituussuhde kasvaa ajan myötä, se voi tarkoittaa, että käytetty kieli on muuttunut monimutkaisemmaksi tai sisältää enemmän taivutettuja sanoja. Jos pituussuhde sen sijaan pienenee, kieli on voinut yksinkertaistua tai muuttua puhekielisemmäksi.

4. TULOKSET JA NIIDEN TARKASTELU

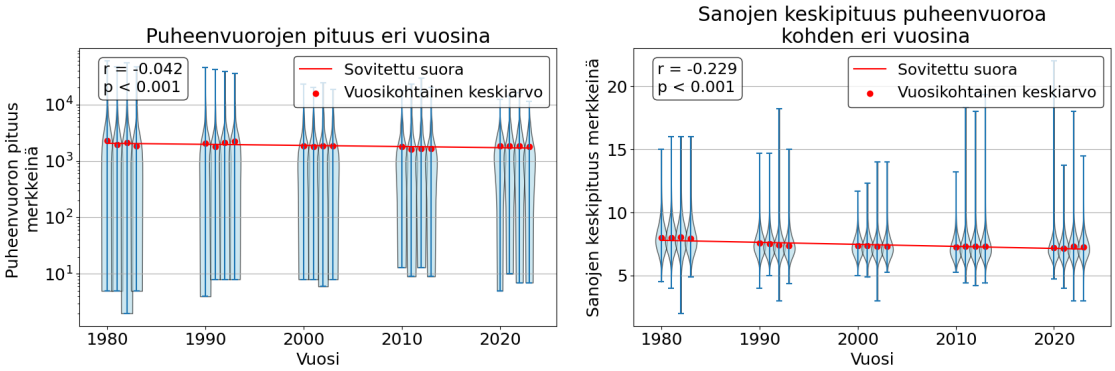
Jokaisen analysoidun ominaisuuden muutoksia 80-luvun ja nykypäivän välillä havainnollistetaan kuvaajien avulla, joissa x-akselilla esitetään vuosiluvut ja y-akselilla kyseiseen ominaisuuteen liittyviä laskettuja tunnus- tai suhdelukuja. Kuvissa on käytetty viuludiagrammeja (violin plot), joiden symmetrisistä tiheyskäyristä voidaan tulkita arvojen jakautumista. Viuludiagrammit ovat saaneet nimensä viulua muistuttavasta muodostaan: mitä leveämpi kohta ”viulussa” on, sitä enemmän arvoja kyseisellä kohdalla on.

Kuhunkin kuvaajaan on sovitettu suora lineaariregressiolla, ja tutkittavan ominaisuuden ja vuosilukujen välistä korrelaatiota kuvataan Pearson-korrelaation r avulla, joka on nähtävissä kunkin kuvan vasemmassa yläkulmassa yhdessä korrelaatioanalyysin p-arvon kanssa. Jokaiseen kuvaan on myös merkattu ominaisuuteen liittyvien arvojen keskiarvo vuosikohtaisesti. Kaikki tuloskuvaajat ovat nähtävissä kuvissa 4.1–4.5.

4.1 Tulokset

Kuvan 4.1 perusteella huomataan, että puheenvuorojen pituudessa ei ole tapahtunut suurta muutosta ajan saatossa. Kuvaajan perusteella puheenvuorojen pituus on pysynyt keskimäärin noin 2 000–2 300 merkin kohdalla aina 80-luvulta nykypäivään. Pearsonin korrelaatiokertoimen arvo $r = -0,042$ ($p < 0,001$) kertoo erittäin heikosta mutta tilastollisesti merkittävästä negatiivisesta korrelaatiosta. Yksittäisten puheenvuorojen pituuksissa taas on suurta vaihtelua. Lyhimmät puheet ovat muutamien merkkien pituisia, kun taas pisimmät lähestyvät 60 000 merkin pituutta. Analyysistä on rajattu pois 60 000 merkin ylittävät puheenvuorot poikkeavuuksina. Käytännössä poikkeuksellisen pitkiä puheenvuoroja oli koko noin 150 000 puheenvuoron aineistossa vain yksi.

Sanojen keskipituudessa on sen sijaan havaittavissa ajallinen muutos kuvan 4.1 perusteella. Korrelaatio $r = -0,229$ ($p < 0,001$) kertoo negatiivisesta riippuvuudesta sanapituuden ja vuoden välillä, eli vuosiluvun kasvaessa keskimääräinen sanapituus lyhenee. Kyseessä ei ole erityisen vahva korrelaatio, joskin se on tilastollisesti merkittävä. 1980-luvun alussa keskimääräinen sanapituus on kuvaajan perusteella noin kahdeksan merkkiä. 2020-luvulle tultaessa se on noin seitsemän merkkiä, eli muutos on yhden merkin luokkaa tarkasteluvälin alun ja lopun välillä. Lyhentynyt sanapituus voi viitata kielen yksinkertaistumiseen tai lyhyempien sanojen käyttöön.

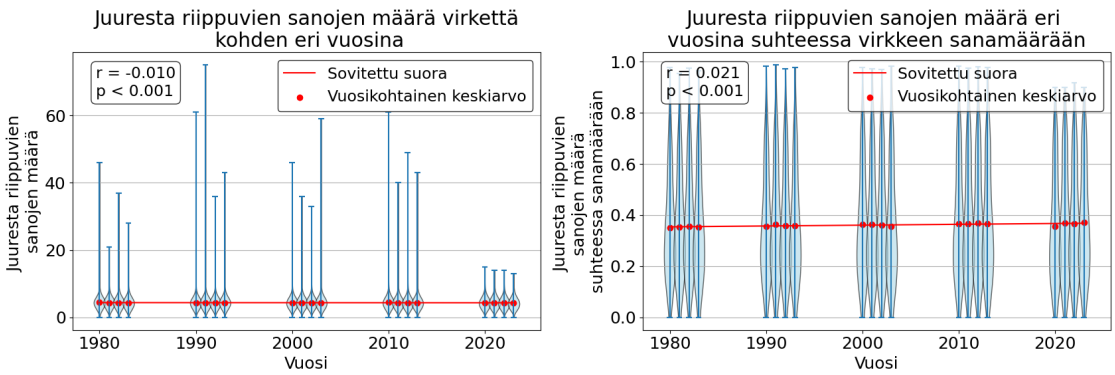


Kuva 4.1. *Puheenvuorojen pituuksien ja sanojen keskipituuksien ajallisia muutoksia havainnollistavat kuvaajat. Puheenvuorojen pituus merkkeinä esitetään logaritmisella asteikolla, sillä sen tulkitseminen on tässä tapauksessa selkeämpää. Yli 60 000 merkkiä ylittäviä puheenvuoroja pidetään poikkeavuuksina ja ne on rajattu pois analyysistä.*

Yleisellä tasolla 7–8 merkin sanapituus on jossain määrin linjassa suomen kielen sanapituuksien aikaisemman tutkimuksen kanssa. Heikkinen et al. (2001) tutki muun muassa suomen kielen sanapituuksia niin sanotusta Parole-korpuksesta, joka sisältää noin 21 miljoonaa sanaa. Aineisto on kerätty lähinnä 90-luvun teksteistä eri lähteistä, kuten sanomalehdistä, tietokirjoista, romaaneista ja aikakauslehdistä (Heikkinen et al. 2001). Koko korpuksen pohjalta laskettu keskimääräinen sanapituus oli 8,5 merkkiä. Oulun korpuksena tunnetusta 60-luvun aineistosta laskettu keskimääräinen sanapituus oli 7,5 merkkiä (Niemikorpi 1996, Heikkinen et al. 2001 mukaan). Parlamenttiaineistolla laajennetusta Oulun korpuksesta laskettuna keskimääräiseksi sanapituudeksi saatiin 7,4 grafeemia (Pääkkönen 1990, Heikkinen et al. 2001 mukaan). Seitsemän merkin keskimääräinen sanapituus on siis lyhyehkö, mutta sitä voi selittää muun muassa se, että puheenvuorotranskriptiot ovat puheenomaisempaa kieltä kuin moni muu kirjoitettu aineisto.

Kuvasta 4.2 nähdään, että keskimääräisesti juuresta riippuvien sanojen määrässä ei ole käytännössä tapahtunut muutosta ajan saatossa. Korrelaatiokerroin $r = -0,010$ $(p < 0,001)$ on hyvin lähellä nollaa, eli vaikka tilastollisesti merkittävä yhteys löytyykin, merkittävää vaikutusta ajankohdalla ei ole juuresta riippuvien sanojen määrään. Tämä pätee myös sanamäärään suhteutetulle juuresta riippuvien sanojen määrälle. Virkkeiden syntaktinen kompleksisuus on pysynyt siis jossain määrin samalla tasolla. Mittarien perusteella eduskunnan puheenvuorojen kieli on kuitenkin melko kompleksista itsessään, sillä juuresta riippuvien sanojen määrä on suurehko, joka viittaa useisiin alisteisiin rakenteisiin virkkeissä.

Kuvan 4.2 perusteella juuresta riippuvien sanojen määrä on ollut keskimäärin noin viisi koko tarkasteluajanjakson aikana. Monissa virkkeissä juuresta riippuvien sanojen määrä kipuaa kuitenkin lähemmäs kymmentä ja jopa kymmeniin. Eduskunnan puheenvuoroissa esiintyy siis hyvin monimutkaisiakin lauserakenteita. Tätä voi selittää puheenvuorojen tie-



Kuva 4.2. *Juuresta riippuvien sanojen määrän kehitykseen liittyvät kuvaajat. Vasemmanpuoleisessa kuvaajassa on esitetty juuresta riippuvien sanojen virkekohtaisen määrän analyysi. Oikeanpuoleisessa kuvassa juuresta riippuvien sanojen määrät on lisäksi suhteutettu virkkeen sanamäärään.*

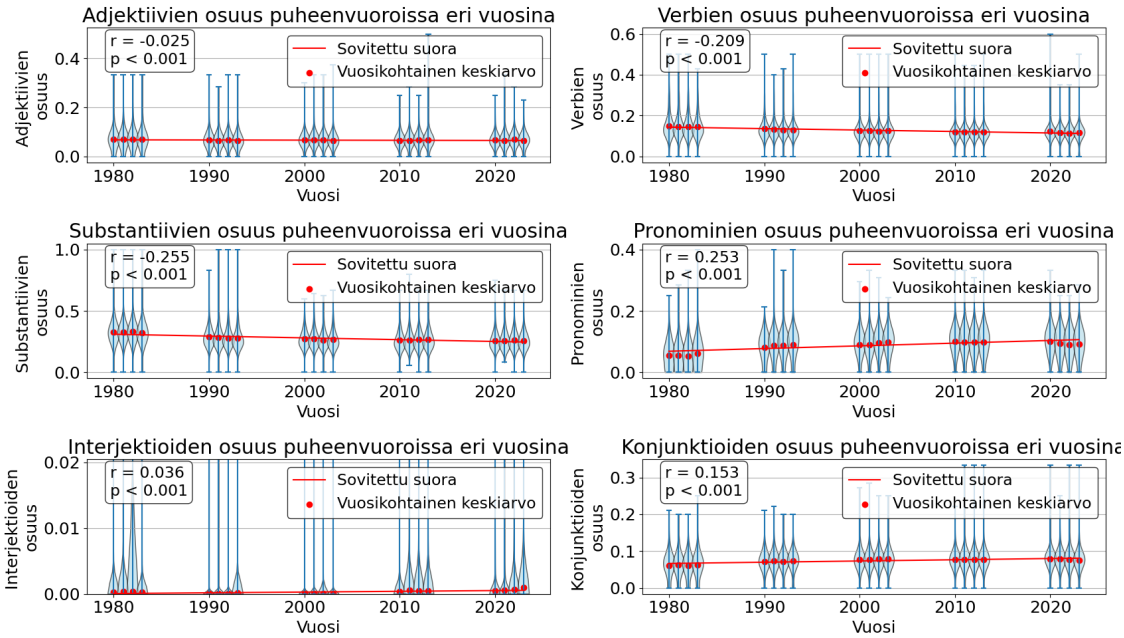
tynlainen puheenomaisuus. Puhuttu kieli sallii joustavampia ja pidempiä lauserakenteita kuin kirjoitettu teksti. Juuresta riippuvien sanojen määrä voi nousta esimerkiksi silloinkin, kun virkkeet pitenevät listausten tai tarkennusten vuoksi.

Sanaluokkien käytössä eduskunnan puheenvuoroissa on tapahtunut joitakin muutoksia viime vuosikymmenten aikana. Eri sanaluokkien osuuksien kuvaajat ovat eriteltynä kuvassa 4.3. Kuvasta nähdään, että adjektiivien osuus puheenvuoroissa ei ole keskimäärin juurikaan muuttunut tarkasteltavien vuosien aikana. Niiden osuus on pysytellyt keskimäärin noin 8–9% pinnassa. Korrelaation perusteella adjektiivien osuus ei myöskään juurikaan muutu ajan myötä, vaikka erittäin pieni tilastollinen riippuvuus niillä onkin ($r = -0,025$, $p < 0,001$).

Verbien osuudessa sen sijaan on tapahtunut merkittävämpi muutos. Kuvaajan perusteella 80-luvun alun puheenvuoroissa verbien osuus on ollut noin 15% puheenvuoron sanoista. 2020-luvulla lukema on 11% luokkaa, eli verbien osuus on laskenut noin neljä prosenttiyksikköä. Myös korrelaatiokerroin $r = -0,209$ ($p < 0,001$) viittaa siihen, että vuosiluvun kasvaessa verbien osuus laskee.

Substantiivien osuus eduskunnan puheenvuoroissa on vähentynyt ajan saatossa. Kuvaajan perusteella substantiivien osuus on ollut tarkasteluvälin alkutilanteessa noin 32% ja lopputilanteessa noin 26%, eli muutos on kuuden prosenttiyksikön luokkaa. Tämä on huomattava muutos ja sitä tukee myös korrelaatiokertoimen arvo $r = -0,255$ ($p < 0,001$). Vuoden ja substantiivien osuuden välillä on negatiivinen korrelaatio, eli vuosiluvun kasvaessa substantiivien osuus puheenvuoroissa pienenee.

Vastaavasti pronominien sekä konjunktioiden osuudet aineiston puheenvuoroissa kasvavat. Erityisesti pronominien osalta kyse on merkittävästä muutoksesta. 1980-luvun alun puheenvuoroissa pronominien osuus on noin kuuden prosentin tasolla ja 2020-luvun alussa noin 11% luokkaa. Tällöin on pronominien osuus puheenvuoroista kasvanut jopa viidel-



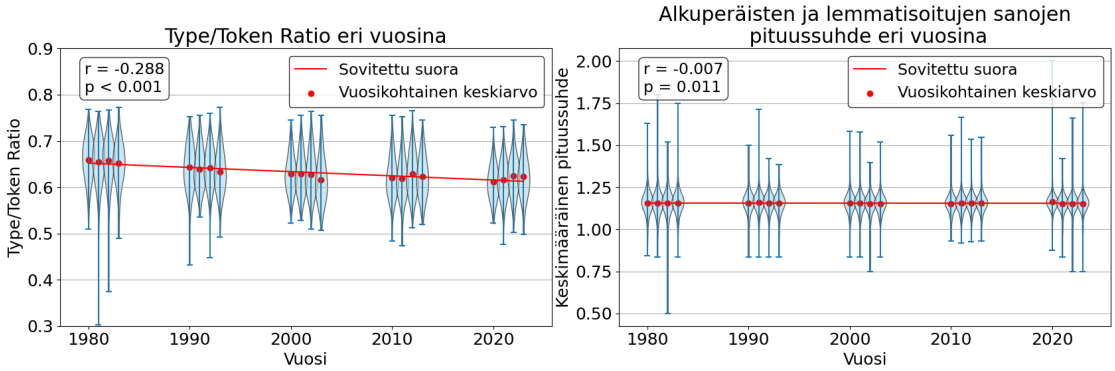
Kuva 4.3. *Eri sanaluokkien osuuksia kuvaavat tuloskuvaajat. Y-akseli kuvaa kunkin sanaluokan suhteellista osuutta (0–1) puheenvuoron sanoista ja x-akselilla on vuosiluvut. Huomioitavaa on, että verbien osuuteen ei ole laskettu mukaan apuverbejä ja substantiivien osuudessa ei ole huomioitu erisnimiä. Interjektioiden osalta y-akseli on rajattu 0–2% välille havainnollistuksen vuoksi, sillä kahden prosentin yläpuolelle sijoittuu lähinnä poikkeusarvoja.*

lä prosentilla. Konjunktioiden osalta muutos ei ole yhtä huomattava, kuvaajan perusteella noin kahden prosenttiyksikön luokkaa ($6\% \rightarrow 8\%$).

Kuvan 4.3 perusteella interjektioiden osuus on äärimmäisen matalalla eikä niiden osuuksissa puheenvuoroista ole havaittavissa juurikaan muutosta. Interjektioiden kuvaajasta nähdään, että niiden käyttö on eduskunnan puheenvuoroissa hyvin vähäistä, sillä keskimääräinen osuus puheenvuoroissa on lähellä nollaa prosenttia.

Sanaston monimuotoisuutta tutkittiin hyödyntäen TTR-lukuja. Tämän analyysin tulokset näkyvät vasemmanpuoleisesta kuvaajasta kuvassa 4.4. Korrelaatio $r = -0,288$ ($p < 0,001$) viittaa lineaariseen yhteyteen: vuosiluvun kasvaessa TTR pienenee. 1980-luvun alussa keskimääräinen TTR on kuvaajan perusteella noin $0,66$ ja 2020-luvun alussa taas noin $0,62$, eli muutos on suurin piirtein $0,04$ suuruinen. Kuvaajasta voidaan tulkita eduskunnan puheenvuoroissa käytetyn sanaston olevan kohtalaisen monipuolista, sillä keskimääräiset TTR-lukemat on melko korkeita. Kokonaisuudessaan voidaan siis todeta, että kielen variaatio on vähentynyt ajan myötä, mutta kieli on itsessään siitä huolimatta monipuolista.

Kuvassa 4.4 oikealla on esitetty kuvaaja alkuperäisten ja lemmatisoitujen sanojen pituussuhteen kehityksestä eri vuosina. Keskimääräinen pituussuhde on pysytellyt noin $1,15$ kohdalla koko tarkasteluajanjakson ajan. Tämä tarkoittaa, että alkuperäiset sanat ovat



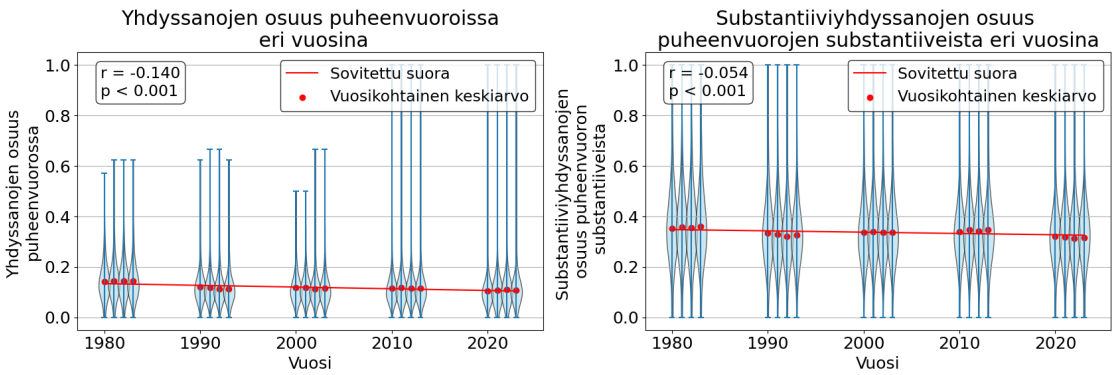
Kuva 4.4. *Vasemmalla on sanaston monimuotoisuuden mittarin TTR-luvun (type/token ratio) analyysin tuloskuvaaja, ja oikealla sanan ja sen lemmatisoidun vastineen pituussuhteen tutkimuksen tuloskuvaaja.*

noin 15% lemmatisoituja sanoja pidempiä, mikä oli odotettava tulos. Ajallista muutosta ominaisuuden arvoissa ei ole tapahtunut, mihin viittaa myös hyvin lähellä nollaa oleva korrelaatiokerroin $r = -0,007$ ($p = 0,011$). Vakiintunut pituussuhde kertoo siitä, että taivutusmuotojen käytössä ei ole tapahtunut käytännön muutosta.

Yhdyssanojen osuuksien analyysien tuloskuvaajat on esitetty kuvassa 4.5. Yhdyssanojen osuuksia puheenvuoroissa tarkasteltiin kahdesta näkökulmasta: kuinka suuri osuus puheenvuoron sanoista on yhdyssanoja (vasen kuvaaja), ja kuinka suuri osa puheenvuoron substantiiveista on yhdyssanoja (oikea kuvaaja). Vasemmanpuoleisesta kuvaajasta nähdään, että yhdyssanojen osuus puheenvuoroissa on vähentynyt ajan kanssa. Muutos on suurin piirtein kolmen prosenttiyksikön luokkaa (noin $14\% \rightarrow 11\%$) eli kyse ei ole valtavasta muutoksesta, joskin se on silti tilastollisesti merkittävää. Vuosiluvun ja yhdyssanojen osuuden välillä on heikko negatiivinen korrelaatio ($r = -0,140$, $p < 0,001$), eli vuoden kasvaessa yhdyssanojen osuus pienenee.

Oikeanpuoleisen kuvaajan perusteella substantiiviyhdyssanojen osuuksissa puheenvuorojen substantiiveista ei ole kuitenkaan suurta muutosta. Korrelaatiokerrointa $r = -0,054$ tulkitsemalla nähdään, että vuoden ja substantiiviyhdyssanojen osuuksien välillä on vain erittäin heikko negatiivinen korrelaatio. Osuus on pysynyt keskimäärin noin 33% ja 35% välillä tarkasteluvuosien aikana.

Yhdyssanojen osuus puheenvuorojen substantiiveista ei ole kokenut huomattavaa ajallista muutosta, joten yhdyssanojen osuuden väheneminen puheenvuoroissa johtuu muista syistä. Koska yleisesti suomen kielessä merkittävä osa yhdyssanoista on substantiiveja, tulokset viittaavat siihen, että yhdyssanojen osuus puheenvuoroissa on pienentynyt substantiivien vähenemisen myötä.



Kuva 4.5. *Kuvaajat yhdyssanojen osuuksien analyysin tueksi. Vasemmalla kuvaaja, jonka avulla on tutkittu yhdyssanojen osuuksia puheenvuoroista ja oikealla kuvaaja yhdyssanojen osuuksien muutoksista puheenvuorojen substantiiveista.*

4.2 Pohdintaa

Työssä mitattujen eri ominaisuuksien analyysien tulokset heijastavat tiettyjä toisiinsa liittyviä ilmiöitä. Ensinnäkin substantiivien käytön suhteellinen väheneminen on todennäköisesti yhteydessä pronominien suhteellisen käytön lisääntymiseen eduskunnassa pidetyissä puheenvuoroissa. Pronomineja voidaan käyttää substantiivi-ilmausten sijasta kun tekstissä tai puheessa viitataan esimerkiksi ihmiseen, eläimeen tai asiaan (*Kielitoimiston ohjepankki* n.d.[b]). On mahdollista, että uudemmissa puheissa asioita kuvataan vähemmän niiden oikeilla nimillä (substantiiveilla) ja sen sijaan käytetään pronomineja, kuten vaikkapa ”se”, ”minä” tai ”hän”. Tämän tapainen muutos voisi viitata myös puheenvuorojen yksinkertaistumiseen tai puhekielistymiseen.

Eri pronomineja on huomattavasti vähemmän kuin erilaisia substantiiveja. Tämä voi olla yhteydessä sanaston vähentyneeseen variaatioon: käytetty sanasto yksipuolistuu kun uniikkeja substantiiveja käytetään suhteessa vähemmän ja samat pronominit sen sijaan toistuvat.

Verbit kuvaavat yleensä toimintaa tai tekemistä. Verbien suhteellisen käytön vähentyminen voi kertoa dynaamisen kielenkäytön vähenemisestä. Esimerkiksi kuvailua, analysointia, esittelyä tai raportointia voi olla uudemmissa puheenvuoroissa enemmän aktiivista toimintaa kuvaavien sanojen sijaan. Konjunktioiden suhteellisen osuuden kasvu taas voi kertoa esimerkiksi siitä, että syy- ja seuraussuhteita tai vastakkainasetteluja ilmaistaan konjunktioiden (kuten ”koska”, ”jos” tai ”vaikka”) avulla uudemmissa puheenvuoroissa enemmän kuin vanhemmissa.

Puheenvuorojen lauserakenteiden kompleksisuudessa ei ole tapahtunut suurta ajallista muutosta viitaten juuresta riippuvien sanojen määrään virkettä kohden. Sen sijaan muutokset vaikuttavat koskevan enemmän sanavalintoja. Sanojen keskipituus on lyhentynyt, sanaluokkien osuudet ovat kokeneet muutoksia ja sanaston monipuolisuus on vähenty-

nyt. Keskimäärin lyhyempien sanojen käyttö ja kielen variaation väheneminen voivat myös viitata puhekielisempään kielenkäyttöön nykypäivän eduskunnan täysistuntojen puheenvuoroissa. Lisäksi joitakin sanoja ja ilmauksia on voitu korvata lyhenteillä niiden yleistyessä (esimerkiksi Euroopan unioni → EU, Yhdistyneet kansakunnat → YK, Yhdysvallat → USA).

4.3 Tutkimuksen rajoitteet

Tähän tutkimukseen liittyy jonkin verran käytännön haasteita, jotka on hyvä ottaa huomioon. Ensinnäkin tutkimusaineisto on haasteellinen. Koska aineiston puheenvuorot ovat transkriptioita eduskunnan puheenvuoroista, ne eivät vastaa täysin alkuperäisiä puhuttuja puheenvuoroja. Transkriptiot ovat kirjaajan referoimia. Vaikka kirjaaminen tapahtuu erityisesti nykyään systemaattisesti tiettyjen yksityiskohtaisten ohjeiden mukaan, vuosikymmenten aikana kirjaamiskäytännöissä on todennäköisesti pientä vaihtelevuutta. Myös yksittäisten kirjurien tekemät tilannekohtaiset valinnat vaikuttavat transkriptioiden eroavuuteen alkuperäisestä puheesta.

Kuitenkin, transkriptiot ovat toistaiseksi ainoa aineisto, joka mahdollistaa eduskunnan kielen tutkimuksen usean vuosikymmenen kattavasta aineistosta. Kielen pitkän aikavälin muutosten tutkimusta ei voida perustaa pelkästään videoitujen täysistuntolähetysten varaan, sillä niitä on saatavilla vasta vuodesta 2008 alkaen. Voidaan myös olettaa, että lähtökohtaisesti kirjaamisen yhteydessä tehdyt suuret muutokset eivät kohdistu puheiden sanavalintoihin tai lauseiden syntaktisiin rakenteisiin, ja kirjaamiskäytännöt ovat riittävän systemaattisia koko tarkasteluajanjakson 1980–2023 ajalta tietyn tyyppisten kielenkäytön ajallisten trendien havaitsemiseksi.

Puheiden tutkimisessa transkriptioiden kautta piilee kaikesta huolimatta myös eräs etu. Yksittäisten kansanedustajien välillä on puhujakohtaisia eroavaisuuksia, joilla voi olla pieni vaikutus tutkimustuloksiin. Puheenvuorojen transkriptoinnin voidaan kuitenkin katsoa tasoittavan joitakin puhujakohtaisia eroja, joka on etu tutkittaessa yleisesti kielenkäytön muutoksia puhujakohtaisten eroavaisuuksien sijaan.

Aineistossa on erittäin monien eri puhujien pitämiä puheenvuoroja. Henkilökohtaiseen puhumistyyliin voi kuulua erilaisia täytesanoja tai änkytyksiä, jotka kertovat enemmän henkilön esiintymistavasta kuin yleisestä kielenkäytöstä eduskuntaympäristössä. Tämänkaltaisia puheen osia kuitenkin editoidaan puheenvuoroista pois kirjaamisen yhteydessä, jolloin pääpaino kirjatuissa puheissa jää sanavalinnoille, puheen sisällölle ja lauserakenteille.

Transkriptioissa on jonkin verran ”kirjoitusvirheitä” ja silmämääräisesti niitä vaikuttaa olevan enemmän vanhemmissa transkriptioissa. Käytännössä virheet ovat niin sanottuja lukuvirheitä, jotka ovat tapahtuneet muunnettaessa vanhoja PDF-muotoisia pöytäkirjoja

tekstiksi tekstintunnistuksen keinoin (optical character recognition, OCR). Jotkin kirjoitusvirheet ovat melko systemaattisia ja toistuvia, esimerkiksi 80-luvun transkriptioissa sana ”herra” on tunnistettu toistuvasti sanaksi ”hetta”. Jotkin virheet taas korjaantuvat erikoismerkkien poiston myötä. Kirjoitusvirheillä voi olla pieni vaikutus tutkimustuloksiin, joskin virheellisten sanojen osuus noin 150 000 puheenvuoron joukossa on kuitenkin pieni.

Aineistossa on joitakin lisäkommentteja puheenvuorojen seassa. Nämä voivat kuvata salin tapahtumia, kuten välihuutoja tai tekoja. Puheenvuorotekstin keskellä voi olla esimerkiksi kohta ”(puhemies koputtaa)”. Näitä lisäkommentteja ei pyritty poistamaan aineistosta, sillä niiden ei oleteta vaikuttavan tuloksiin. Lisäkommenttien huomioimisella olisi suurempi merkitys tehtäessä puhujakohtaisia analyysejä. Kommenttien kirjaamismenetelmät oletetaan siinä määrin samantyylisiksi eri aikakausina, että niitä ei pidetä merkittävinä tutkimuksen tulosten kannalta.

Stanzan NLP-työkalut toimivat suomen kielelle melko luotettavasti, joskin automaattiseen tekstinkäsittelyyn liittyy käytännössä aina virheiden riski. Stanzan lemmatisointi-työkalu tekee joitakin virheitä: sanan perusmuotoa ei määritetä oikein tai yhdyssanojen osasanoja ei tunnisteta oikein, jolloin ”#”-merkkejä ei sijoiteta aina oikeisiin kohtiin. Sanaluokkien merkkauksessa voi myös tapahtua virheitä. Stanzan oletetaan kuitenkin tekevän samoja virheitä järjestelmällisesti eri vuosien puheenvuoroille. Tällöin sen vaikutus ajallisten muutosten tutkimukselle on todennäköisesti vähäinen. Virheiden systemaattisuus ei vääristä eri vuosien välistä vertailua samalla tavalla kuin satunnaiset virheet. Koska tutkimus perustuu suureen datamäärään, eivät yksittäiset virheelliset lemmat tai sanaluokkavalinnat vaikuta merkittävästi kokonaiskuvaan. Automaattisen analyysin rajoitteet ovat silti tärkeitä tiedostaa ja huomioida.

Lisäksi, kielensuodatus käyttäen langdetect-kirjastoa ei ole täysin ongelmatonta. On todennäköistä, että langdetect ei luokittele jokaista sanaa oikeaan kielikategoriaan. Etenkin hyvin lyhyet sanat ovat alttiita väärälle kieliluokitukselle. Kirjoitusvirheet tekstissä ja harvinaisemmat sanat voivat myös johtaa vääriin luokituksiin.

Kielensuodatuksessa käytettiin menetelmää, jossa poistettiin aineistosta puheenvuorot joissa tunnistettiin olevan tiettyä raja-arvoa suurempi osuus jotakin muuta kieltä kuin suomea. Tämä pienentää yksittäisten sanojen väärän luokittelun mahdollista vaikutusta tutkimustuloksiin. Lisäksi käytetty menetelmä pienentää huomattavasti sitä riskiä, että tutkimukseen tulisikin mukaan esimerkiksi ruotsinkielisiä puheenvuoroja, joskin myös suomenkielisiä puheenvuoroja voi poistua. Tällä ei kuitenkaan ole merkittävää vaikutusta kokonaiskuvaan, sillä aineistoa on hyvin paljon käytössä siitäkin huolimatta, että osa suomenkielisistäkin puheenvuoroista jäisi pois tutkimuksesta.

Jatkossa tutkimuksen kannalta olisi hyvä manuaalisesti tarkastella Stanzan tekemän käsittelyn sekä langdetectin tekemän kieliluokittelun tulosten oikeellisuutta. Esimerkiksi lemmatisoinnin, sanaluokkien merkkauksen ja kieliluokittelun tulosten oikeellisuuden tarkis-

taminen manuaalisesti olisi jatkotutkimuksen kannalta tarpeellinen toimenpide. Kandidaatintyön laajuuden huomioiden tämä vaihe kuitenkin rajattiin tämän tutkimuksen ulkopuolelle. Silmämääräisen arvion mukaan kielensuodatus sekä Stanzan avulla tehty tekstinkäsittely ovat hyvin onnistuneita ja palvelevat tutkimuksen tavoitteita löytää kielen ajallisia muutoksia.

4.4 Eduskunnan kielenkäytön tutkimus jatkossa

Parlamenttipuheen ja kielentutkimus on aina ajankohtaista. Nyt kun parlamenttiaineistojen automaattinen analyysi ja keskinäinen vertailu on enenevissä määrin mahdollista ja saavutettavaa, on tutkimusmahdollisuuksien määrä valtava. Mahdollisia ja perusteltuja tutkittavia kielen ominaisuuksia olisi tässä tutkimuksessa tutkittujen lisäksi runsaasti lisää. Eduskuntapuheen ajallisia muutoksia tarkasteltaessa voitaisiin tutkia lisäksi esimerkiksi tiettyjen sanojen esiintyvyyttä puheenvuoroissa tai kielen kompleksisuuden kannalta riippuvuuspuiden maksimisyvyyksiä.

Tässä tutkimuksessa eduskunnan täysistuntojen kieltä tarkasteltiin yleisemmällä tasolla, mutta tutkimusta voisi laajentaa myös tarkemmille tarkastelutasoille. Tämän tutkimuksen pohjalta voitaisiin muun muassa vertailla puolueiden osuuksia kielen muutoksiin tai tarkastella yksittäisten kansanedustajien tai edustajaryhmien kielen kehitystä pitkällä aikavälillä. Lisäksi eduskunnan eri roolien kehitys olisi mahdollinen tutkimuskohde. Kiinnostavaa on esimerkiksi kuinka puhemiehen rooli ja puheenvuorot ovat muuttuneet ajan kanssa, tai kuinka oppositio- tai hallitusrooli vaikuttaa kielenkäytön muutoksiin. Löydettyjen kielen ominaisuuksien ja aikakausikohtaisten erojen pohjalta olisi mahdollista myös kehittää koneoppimiseen perustuva luokitin, jonka avulla voitaisiin ennustaa todennäköinen vuosiluku tai -kymmen yksittäiselle puheenvuorolle.

Se, että aineisto koostuu transkriptioista, vaikuttaa auttamatta aineiston kielen luonteeseen. Transkriptiot eivät ole täydellisiä kopioita alkuperäisestä puheesta. Tämän tutkimuksen kaltaista kielen tutkimusta voitaisiin tehdä myös automaattisen puheentunnistuksen tekemän transkription pohjalta. Tällöin aineisto vastaisi tarkemmin alkuperäistä puhuttua puhetta. Toistaiseksi ongelmana on kuitenkin sellaisen aineiston rajallisuus, josta voitaisiin puheentunnistuksen keinoin luoda laaja eri aikakausia käsittävä transkriptioaineisto.

Jatkotutkimuksissa analyysit kannattaisi toteuttaa kehittyneempiä tilastollisia malleja hyödyntäen, esimerkiksi soveltaen yleistettyjä lineaarisia sekamalleja. Näin analyysissa voitaisiin ottaa samalla huomioon puhujien, puolueiden ja esimerkiksi hallitus- tai oppositioroolien mahdolliset vaikutukset puheenvuorojen ominaisuuksiin tutkittaessa kielen ajallisia muutoksia.

5. YHTEENVETO

Tutkimuksessa selvitettiin, miten suomen kielen käyttö on muuttunut eduskunnan täysistunnoissa vuosien 1980–2023 välillä. Puheenvuoroja analysoitiin valikoitujen kielen ominaisuuksien perusteella (taulukko 3.2) täysistuntojen pöytäkirjoista. Valittujen ominaisuuksien avulla saatiin selville tietoa muun muassa puheenvuorojen sanavalinnoista, sanaston monimuotoisuudesta ja kielen syntaktisesta kompleksisuudesta ajan saatossa. Jokaisesta puheenvuorosta laskettiin joukko kielen rakennetta kuvaavia tunnuslukuja, joita verrattiin vuosilukuihin korrelaatioanalyysin keinoin ajallisten muutosten selvittämiseksi.

Analyysi perustui Parlamenttisampo.fi-portaalin kautta laajaan eduskuntapuheenvuorojen aineistoon. Aineistosta suodatettiin muut kuin suomenkieliset puheenvuorot pois, jonka jälkeen aineisto sisälsi yhteensä noin 150 000 puheenvuoroa 20 eri vuodelta. Lisäksi puheenvuorot käsiteltiin Stanza-analyysipaketin luonnollisen kielen käsittelyn työkaluilla: puheenvuorot tokenisoitiin ja lemmatisoitiin, selvitettiin sanojen sanaluokat sekä niiden väliset riippuvuussuhteet virkkeissä. Esikäsitellyn aineiston pohjalta oli mahdollista laskea lingvistisiä ominaisuuksia kuvaavia tunnuslukuja automaattisesti suuresta aineistosta.

Tutkimuksen tulokset kertovat erityisesti hienoisista sanaston muutoksista: keskimääräinen sanapituus on lyhentynyt, ja substantiivien sekä verbien osuus puheenvuoroissa on pienentynyt. Samalla pronominien ja konjunktioiden osuus on kasvanut. Adjektiivien ja interjektioiden osuudet puheenvuoroissa ovat sen sijaan pysyneet keskimäärin muuttumattomina tarkasteluajanjakson aikana.

TTR-lukemien lasku viittaa kielen variaation hienovaraiseen vähenemiseen. Uudemmissa eduskunnan puheenvuoroissa siis käytetään keskimäärin suppeampaa sanastoa. Yhdyssanojen osuus puheenvuoroissa on ajan myötä vähentynyt, mutta tämä vaikuttaa liittyvän suoraan substantiivien vähenemiseen. Eduskunnan puheenvuorojen kielen rakenteiden syntaktisessa kompleksisuudessa ei tulosten perusteella ole ajan saatossa tapahtunut huomattavaa muutosta.

Tämä tutkimus tarjoaa lähtökohdan kieliteknologiselle ja poliittiselle jatkotutkimukselle. Analyysimenetelmiä voidaan soveltaa tarkemman tarkastelutason tutkimuksiin, kuten yksittäisten puolueiden, kansanedustajien tai kansanedustajaryhmien kielen muutosten tutkimukseen. Lisäksi menetelmiä voidaan soveltaa laajempien kielen muutosten tutkimuk-

siin suurista tekstikorpuksista. Tämä työ tarjoaa uutta tietoa poliittisen kielen ajallisista muutoksista ja osoittaa, kuinka teknologiset menetelmät voivat tukea kielen ajallisten muutosten ja politiikan tutkimusta.

Tässä työssä käytetyt Python-kieliset ohjelmakoodit ovat vapaasti saatavilla osoitteesta `https://github.com/veeranurminen/parliament_language_processing`. Koodit kattavat aineiston esikäsittelyn, kielellisten piirteiden laskemisen sekä analyysiin ja visualisointeihin liittyvät toiminnallisuudet.

LÄHTEET

Akhmetov, I., Pak, A., Ualiyeva, I. ja Gelbukh, A. (2020). Highly language-independent word lemmatization using a machine-learning classifier. *Computacion y Sistemas* 24.3, s. 1353–1364.

Berman, J. J. (2016). *Data simplification: taming information with open source tools*. 1st edition. Elsevier. Luku 4.

Bird, S., Klein, E. ja Loper, E. (2009). *Natural Language Processing with Python*. 1st edition. O'Reilly Media, Inc.

CLARIN in a Nutshell (n.d.). URL: https://www.clarin.eu/content/clarin-nutshell (viitattu 11. 02. 2025).

Danilák, M. (2014). *langdetect*. URL: https://github.com/Mimino666/langdetect (viitattu 14. 04. 2025).

de Marneffe, M.-C., Manning, C. D., Nivre, J. ja Zeman, D. (2021). Universal Dependencies. *Computational linguistics - Association for Computational Linguistics* 47.2, s. 255–308.

Dozat, T. ja Manning, C. D. (2017). *Deep Biaffine Attention for Neural Dependency Parsing*. arXiv: 1611.01734 [cs.CL].

Eduskunnan työjärjestys 40/2000 50–52 § (2000). URL: http://data.finlex.fi/eli/sd/2000/40/ajantasa/2023-02-22/fin.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., Does, J. de, Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., Macedo, L. D. de, Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Darǵis, R., Ring, O., Heusden, R. van, Marx, M. ja Fišer, D. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation* 57.1, s. 415–448.

Erjavec, T. ja Pančur, A. (2019). Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings. *TEI Conference and Members Meeting 2019*. Zenodo.

Heikkinen, V., Lehtinen, O. ja Lounela, M. (2001). Kuvia kirjoitetusta suomesta. *Kielikello*. URL: https://kielikello.fi/kuvia-kirjoitetusta-suomesta/ (viitattu 05. 02. 2025).

Hiidenmaa, P. (2005). Näkökulmia yleiskieleen. *Kielikello*. URL: https://kielikello.fi/nakokulmia-yleiskieleen/ (viitattu 05. 11. 2025).

Honnibal, M., Montani, I., Van Landeghem, S. ja Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Hyvönen, E., Sinikallio, L., Leskinen, P., Drobac, S., Leal, R., La Mela, M., Tuominen, J., Poikkimäki, H. ja Rantala, H. (2024). Publishing and Using Parliamentary Linked Data on the Semantic Web: ParliamentSampo System for Parliament of Finland. *Semantic Web*, s. 1–23.

Hyvönen, E., Sinikallio, L., Leskinen, P., Drobac, S., Tuominen, J., Elo, K., La Mela, M., Koho, M., Ikkala, E., Tamper, M., Leal, R. ja Kesäniemi, J. (2021). Parlamenttisampo: eduskunnan aineistojen linkitetyn avoimen datan palvelu ja sen käyttömahdollisuudet. *Informaatiotutkimus* 40, s. 216–244.

Jauhiainen, T., Piitulainen, J., Axelson, E., Dieckmann, U., Lennes, M., Niemi, J., Rueter, J. ja Linden, K. (2024). Investigating Multilinguality in the Plenary Sessions of the Parliament of Finland with Automatic Language Identification. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Toim. D. Fišer, M. Eskevich ja D. Bordon. International conference on computational linguistics. European Language Resources Association (ELRA), s. 48–56.

Johnson, W. (1944). Studies in Language Behavior: A program of research. *Psychological monographs* 56.2, s. 1–15.

Jurafsky, D. ja Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Luku 2,17,19.

Kestilä-Kekkonen, E. (n.d.). Kovarianssi ja korrelaatio. Pearsonin korrelaatiokerroin. *Kvantitatiivisen tutkimuksen verkkokäsikirja*. Tietoarkisto. URL: https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvanti/korrelaatio/korrelaatio/ (viitattu 04. 03. 2025).

Kielitoimiston ohjepankki (n.d.[a]). *Konjunktiot*. URL: https://kielitoimistonohjepankki.fi/ohje/konjunktiot/ (viitattu 30. 04. 2025).

Kielitoimiston ohjepankki (n.d.[b]). *Pronominit: mikä on pronomini?* URL: https://kielitoimistonohjepankki.fi/ohje/pronominit-mika-on-pronomini/ (viitattu 04. 05. 2025).

Lui, M. ja Baldwin, T. (2012). langid.py: An Off-the-shelf Language Identification Tool. *Proceedings of the ACL 2012 System Demonstrations*. Toim. M. Zhang. Association for Computational Linguistics, s. 25–30.

Nakatani, S. (2010). *Language Detection Library for Java*. URL: https://www.slideshare.net/slideshow/language-detection-library-for-java/6014274 (viitattu 14. 04. 2025).

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. ja Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, s. 101–108.

Ranta, A. ja Masciolini, A. (2023). deptreepy. *GitHub repository*. URL: https://github.com/aarneranta/deptreepy.

Rao, P. B., Kotagiri, C., Madamshetti, R., Minkoori, A. ja Gande, S. (2023). Language Detection Using Natural Language Processing. *2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC)*. IEEE, s. 231–237.

Suomen perustuslaki 731/1999 31 § (1999). URL: http://data.finlex.fi/eli/sd/1999/731/ajantasa/2018-10-05/fin.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. ja SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17, s. 261–272.

Voutilainen, E. (2025). Kielellinen huoltovarmuus eduskunnassa. *Kieli, koulutus ja yhteiskunta* 16.2.

Voutilainen, E. (2023). Written representation of spoken interaction in the official parliamentary transcripts of the Finnish Parliament. *Frontiers in Communication* 8. Toim. H. Fraser.

LIITE A: ESIMERKKI EDUSKUNNAN TÄYSISTUNNON PUHEENVUORON TRANSKRIPTIOSTA

Tässä liitteessä on esimerkkipuheenvuoro eduskunnan täysistunnosta vuodelta 2000. Puheenvuoron transkriptio on ladattu Parlamenttisampo.fi-portaalista ja sen on pitänyt kansanedustaja Henrik Lax (Suomen ruotsalainen kansanpuolue) 18.02.2000. Kyseinen puheenvuoro on valittu esimerkiksi, sillä se havainnollistaa hyvin puheiden ja niiden transkriptioiden piirteitä.

Henrik Laxin puheenvuoron transkriptio vuodelta 2000:

> ”Rouva puhemies! Haluaisin vain yhteen asiaan kiinnittää huomiota. Ed. Myllyniemi sanoi, että nyt ei korkeimmille oikeuksille riittänyt omien edustajien nimittäminen lautakuntaan. Tämä lausuma vie ajatukset nyt ihan harhaan. Kysymyshän on siitä, että korkeimmat oikeudet ovat uudistaneet nimityskäytäntönsä. Jotta tämä myös uudistuisi alemmissa oikeusasteissa ja jotta päästään murtamaan se sisäänpäinlämpiävyys ja jäykkyys, mikä siinä on ollut, tämä on se syy, miksi myös muiden tuomioistuinten edustajien valitseminen ilmoittautumisen perusteella on päädytty uskomaan korkeimmille oikeuksille. Se ei ole ollenkaan siitä lähtöisin, mitä korkeimmat oikeudet ovat halunneet.”