TURUN YLIOPISTO

Melkein vai lähes – synonymiaa?

Kieliasiantuntijuuden tutkinto-ohjelma, digitaalinen kielentutkimus
Kieli- ja käännöstieteiden laitos
Humanistinen tiedekunta
Turun yliopisto
Pro gradu -tutkielma

Laatija:
Petra Lähde

30.12.2023
Turku

Pro gradu -tutkielma

Oppiaine: Digitaalinen kielentutkimus
Tekijä(t): Petra Lähde
Otsikko: Melkein vai lähes – synonymiaa?
Ohjaaja(t): professori Veronika Laippala
Sivumäärä: 45
Päivämäärä: 30.12.2023

Ihmiskielissä on samanmerkityksisiä sanoja, joita voidaan kutsua synonyymeiksi. Kaikki samatarkoitteiset sanat eivät kuitenkaan jaa keskenään samaa merkityskenttää täydellisesti. Koska täydellinen synonymia on harvinaista, on syytä olettaa, että läheisilläkin synonyymeilla on joitakin eroja, jolloin voidaan puhua melkein-synonymiasta.

Tässä kognitiiviseen kielioppiteoriaan pohjaavassa tutkimuksessa käsitellään suomen kielen astemääritteitä *melkein* ja *lähes* adjektiivien yhteydessä. Lähtökohta on, että sanojen välille voidaan löytää eroavaisuuksia, ja siten sanat ovat melkein-synonyymejä. Tämän tutkimiseen käytettiin riippuvuussuhdeprofiileja, jotka perustuvat sanojen syntaktisiin riippuvuussuhteisiin spontaanissa kielenkäytössä. Riippuvuussuhdeprofiilit eivät vaadi suuria teoreettisia pohjaolettamia, vaan ne muodostuvat yksinkertaisesti sanojen syntaktisista suhteista.

Tutkimuksen aineisto koostuu Finnish Internet Parsebankin niistä virkkeistä, jotka sisältävät joko sanan *lähes* tai *melkein* adjektiivin yhteydessä. Aineisto on jäsennetty syntaktisesti noudattaen Universal Dependencies –skeemaa. Jäsennellystä aineistosta muodostettiin biarceja (kaksoiskaaria), jotka kuvastavat lauseen sisäisiä syntaktisia suhteita. Lopullisesta aineistosta poistettiin kaikki leksikaalinen informaatio. Lauseita oli kaiken kaikkiaan 16 833, joista 8950 sisälsi sanan *melkein* ja 7883 sanan *lähes.*

Tutkimusta varten ohjelmoitiin tukivektorikoneluokittelija, joka tunnistaa lauseen syntaktisten riippuvuussuhteiden perusteella, kumpi tutkimuskohteista tulee lauseessa kyseeseen. Luokittelijan luokat olivat *lähes* ja *melkein,* ja ennustaja lauseiden biarcit. Luokkia tasattiin niin, että kummastakin luokasta poimittiin satunnaisesti 7500 esimerkkiä. Aineisto jaettiin niin ikään satunnaisesti koulutus- ja testidataan (80 % / 20 %), ja mallin koulutus toistettiin 100 kertaa. Mallin oletettu perustaso oli 0.5. Malli saavutti 0.633 osuvuuden (*accuracy*), mikä osoittaa, että sanojen riippuvuussuhdeprofiilit ovat erilaiset.

Luokittelijasta saaduista sanoja erottavista riippuvuussuhdepiirteistä sanalle *lähes* ainoa uniikki piirre olivat *punct,* eli välimerkit; sanalle *melkein* puolestaan *nsubj* eli nominaalisubjekti ja *nmod:poss* eli possessiivinen nominaalimäärite. Molemmissa luokissa esiintyvät *ROOT* (juuri) ja *obl* (obliikvinominaali) osoittivat toisistaan poikkeavia riippuvuusyhtymiä eri luokkien välillä.

Tutkimuksen perusteella näyttäisi siltä, että astemääritteet *melkein* ja *lähes* eivät ole täysin synonyymisia. Lisätutkimusta kuitenkin vaatii se, mitä tällä tutkimuksella esiin saadut riippuvuussuhdepiirteet pitävät sisällään.

Avainsanat: Synonymia, kognitiivinen kielitiede, astemäärite, tukivektorikone, riippuvuussuhdeprofiili, luokittelija, syntaksi

Sisällysluettelo

1 Johdanto

Ihmiskielissä on sanoja, joiden merkitystä voisi pitää ensi silmäyksellä yksioikoisesti synonyymisena: esimerkiksi suomen kielen sanaparit *vain* ja *ainoastaan*, *karhu* ja *otso*, *jänis* ja *pupu* ovat tällaisia samatarkoitteisia sanoja. Nämä kuullessaan kielen puhuja saattaa kuitenkin tuntea hienovaraista ihmetystä. Käyttäisikö hän muka sanaa *pupu* puhuessaan yläkoululaisille suomalaisesta metsästä? Tai miksi sanotaan *vain ja ainoastaan*, jos sanat ovat samanmerkityksisiä? Onko kyseessä vain tarkoitteen intensiteettiä vahvistava toisto, vai kenties kuitenkin asian vahvistaminen useammalta kannalta kahden hieman erimerkityksisen sanan avulla? *Otso*-sanan monet tietävät merkitsevän 'karhua', mutta sanan tyylilaji on erilainen kuin sanan *karhu* – esimerkiksi kielitoimiston sanakirja nimeää sen runokielen sanaksi.

Synonymia tarkoittaa läheisintä semanttista samankaltaisuutta (Miller & Charles 1991; Arppe 2008). Täydellinen synonyymipari olisi tästä päätellen sellainen, jonka osapuolet olisivat joka tilanteessa korvattavissa toisillaan (esim. Cruse 1986). Synonymia on perustavanlaatuinen kielellinen ilmiö, mutta absoluuttinen synonymia on varsin harvinaista (Edmonds & Hirst 2002). Esimerkiksi päältä katsoen synonyymiset *vihta* ja *vasta* ovatkin sinänsä tyyliltään neutraaleja ja ymmärrettävissä täsmälleen saunalätkimisessä käytetyiksi esineiksi, joten miksi ne eivät siis olisi täysin synonyymeja? Niiden käytössä on kuitenkin selkeä alueellinen viittaus – *vihta* on länsi- ja *vasta* itäsuomalainen sana, ja siten niiden keskinäinen vaihtaminen tekstissä ei ole neutraalia (ks. esim. Saukkonen 1984). Tähän tapaan monista synonyymeista voidaan löytää eritasoisia, hyvinkin hienosäikeisiä eroavaisuuksia, jotka osoittavat ne lopulta eritarkoitteisiksi tai eri tyyliä edustaviksi (Edmonds & Hirst 2002). Tämä pro gradu -työ tutkii melkein-synonymiaa, ja tarkemmin sanoen se pyrkii selvittämään suomen kielen sanojen *melkein* ja *lähes* keskinäistä synonymista suhdetta ja mahdollisia eroavaisuuksia, kun ne esiintyvät adjektiivin yhteydessä sen astemääritteenä eli adjektiivin intensiteettiä ilmaisevana etumääritteenä.

Astemääritteiden tehtävänä on ilmaista jonkin ominaisuuden astetta (VISK, Paradis, Huumo). Niiden avulla ominaisuuden asteen esitetään olevan odotuksenmukaista korkeampi tai alhaisempi, esim. ***aika*** *kylmä,* ***melkein*** *kuollut,* ***hemmetin*** *ilkeä* (VISK 2008). *Melkein* ja *lähes* kuuluvat approksimatiivisten astemääritteiden luokkaan, eli ne kuvaavat tilannetta, jossa adjektiivin ilmaiseman ominaisuuden tiettyä rajaa ei ole aivan saavutettu: *melkein tyhjä lasi, lähes täysi katsomo* (Huumo 2021). *Melkein* ja *lähes* esitetään siis usein synonyymisinä sanakirjoissa ja kieliopeissa. Intuitiivisesti sanat vaikuttavat samanmerkityksisiltä.

Esimerkiksi Kielitoimiston sanakirja selittää sanoja toisillaan, joskaan ne eivät ole toistensa täydelliset selitykset (KTS 2023). (VISK §615, Kielitoimiston sanakirja). Kielitoimiston sanakirjasta haettuna *lähes* saa selityksekseen ’melkein, miltei, likipitäen’, *melkein* puolestaan ’vähää vaille, miltei, lähes, liki, likimain’. Sanakirjojen luoma kuva synonymisuudesta voi kuitenkin johtaa harhaan (Biber et al. 1998).

Tämän tutkimuksen perusolettama on, että sanat *melkein* ja *lähes* eivät ole täydellisiä synonyymejä, mikä tarkoittaa, että niiden välillä on oltava eroja (Edmonds & Hirst, 2008). Sanoja *melkein* ja *lähes* olisikin syytä pitää melkein-synonyymeinä (Huumo 2022), jotka siis tuntuvat vastaavaan toisiaan, mutta joilla voidaan kuitenkin havaita erilaisia käyttöä laukaisevia tekijöitä. On myös syytä olettaa, että tietyn sanan käytölle löytyy aiheita sen syntaktis-semanttisesta ympäristöstä (Arppe 2008; Divjak & Gries 2008; Edmods & Hirst 2002) sitä tukevia syitä voidaan hahmottaa tutkimalla sanojen esiintymisympäristön syntaktisia riippuvuussuhteita.

Vertauskuvallisesti voidaan ajatella, että tämän tutkimuksen toiminta-aluetta rajaavat parin *melkein* ja *lähes* sanakirjamääritelmien sisältämät muut sanat, eli se mitä tapahtuu tutkimuskohteiden täsmällisen samankaltaisuuden ulkopuolella: mistä johtuvat sanojen erilaiset kuvaukset, ja millaiset piirteet ja yhteydet erottavat ne toisistaan. Synonyymien voidaan esimerkiksi esittää muodostavan klustereita, joiksi samanmerkityksiset sanat ryhmittyvät sanojen semanttisessa avaruudessa. Samassa klusterissa sijaitsevat sanat jakavat yhteisiä piirteitä, mutta eivät kuitenkaan yhtä merkityskenttää täydelleen (Divjak & Gries 2006). Melkein-synonyymien muodostama klusteri on sisäisesti rakentunut jollakin loogisella tavalla, joka on selitettävissä aiheutuvaksi jostakin niihin vaikuttavasta asiasta (Arppe 2008, 8; ks. myös Divjak & Gries 2008). Muiden sanakirjamääritelmissä esiintyvien sanojen voidaan siis ajatella muovaavan sitä semanttista kenttää, jolla sanojen käyttöyhteydet rakentavat sanojen välisiä merkityseroja. Tämän tutkimuksen fokuksessa ovat semanttisen kontekstin sijaan syntaktiset yhteydet.

Tutkimuksen tavoitteena on sekä tuoda esille sanojen *melkein* ja *lähes* mahdollisia erottavia piirteitä, kun ne esiintyvät adjektiivien yhteydessä, hyödyntäen koneoppimiseen perustuvaa luokittelijaa, että asettua osaksi melkein-synonymian tutkimusta lähestyen tutkittavia kielen aineksia aineistolähtöisesti ja niiden luonnolliseen esiintymiseen perustuen. On myös huomattava, että sanojen välinen ”eroavaisuus” ei, huolimatta sanan merkityksestä, määritä sitä, että eroja on oltava – jos eroja ei esiinnykään, voidaan olettaa samankaltaisuutta. Tutkielman teoreettinen viitekehys on kognitiivisessa kielitieteessä, ja se pohjaa

kognitiiviseen kielioppikäsitykseen, jonka mukaan merkitys ei ole erillään kielenkäyttötilanteesta tai kielen käyttäjästä (Langacker 2008). Tutkimus perustuu siis olettamaan siitä, että sanojen merkitys ei ole irrallaan niiden käyttöympäristöstä (Langacker, 2008), jolloin merkityseroja voivat kuvastaa leksikaalisen esiintymiskontekstin ohella myös syntaktiset riippuvuussuhteet. Tämä on luontevaa, sillä tutkielmassa käytetty aineisto edustaa nimenomaan luonnollisessa, spontaanissa kielenkäytössä esiintyviä esimerkkejä (FIP; Luotolahti et al. 2015).

Tutkimus liittyy myös korpuslingvistiikan alaan, koska aineisto on peräisin suuresta suomalaisesta korpuksesta, joka on koottu suomenkielisestä internetistä, mm. keskustelufoorumeilta ja erilaisilta verkkosivuilta (FIP; Luotolahti et al. 2015). Korpus muodostetaan siis kielenkäyttäjien tuottamasta spontaanista kieliaineistosta (spontaanilla tarkoitetaan tässä sellaista kieltä, jota kielenkäyttäjä tuottaa ilman tutkijan määrittämiä ärsykkeitä tai olosuhteita), ja siten se liittyy myös käyttöpohjaiseen korpustutkimukseen (vrt. Vanhatalo 2003; ks. Ädel 2020 in Gries & Paquot (toim.) 2020). Kielestä muodostetun mallin sijasta siis tutkitaan kieltä sen käyttöyhteyksissä (Ivaska 2015).

Sanojen *melkein* ja *lähes* käytön eroja adjektiivien yhteydessä analysoidaan siis suuren korpusaineiston pohjalta. Tutkimuksissa tarkastellaan käyttöä laukaisevia syntaktisia piirteitä merkityserojen selityksenä. Suuri aineisto tarjoaa mahdollisuuden tutkia synonyymien (tai saman merkitysklusterin sisällä sijaitsevien sanojen) käyttöä laajalla skaalalla, halki tekstilajien ja -tyyppien, sekä mahdollistaa niiden analysoinnin dataorientoituneesti (Laippala et al. 2018; Biber et al. 1998). Tutkimusta vastaavista aiheista on tehty joko nojatuolilingvistisesti, (esim. Huumo, 2022), jolloin työkaluina toimivat kielenkäyttäjän intuitio ja tutkijan kielitaju, tai käsin annotoituun, usein takstityypiltään ja -lajiltaan rajattuun korpukseen pohjautuen (esim. Divjak & Gries, 2006). Nämä lähestymistavat kuitenkin mahdollistavat määrältään ja kattavuudeltaan hyvin rajallisen aineiston tarkastelun.

Keskeisenä työkaluna ja lähtökohtana tutkielmassa toimii riippuvuussuhdeprofiili (*dependency profile*), joka on sanan syntaktisten yhteyksien eli riippuvuuksien muodostama piirteiden kokonaisuus. (Laippala & al. 2018) Riippuvuussuhdeprofiileja voitaisiin luonnehtia sanojen (tai ilmaisujen) sellaisiksi ominaisuuksien kokoelmaksi, jotka syntyvät sanojen kieliopillisesta kontekstista. Näin ollen riippuvuussuhdeprofiilit voivat paljastaa tai selittää sellaisia eroja, joita leksikaalisesti tai semanttisesti lähestyttynä ei ole mahdollista selkeästi havaita.

Tutkimuksen aineisto on peräisin Finnish Internet Parsebankista, joka on koostettu internetistä automaattisesti haetuista teksteistä (*web crawl*) (Luotolahti et al. 2015). Parsebankissa on tällä hetkellä noin 3,7 miljardia sanaa (FIP). Parsebank on valittu tutkielman aineiston lähteeksi, koska sen tarjoama aineisto on hyvin laaja. Se sisältää myös monia tekstilajeja, joten tutkimuksen kohdesanoista saadaan tietoa monenlaisissa käyttöyhteyksissä. FIP:n tarjoamaa korpusta voidaan siis pitää edustavana osana luonnollisesta (kirjoitetusta) suomen kielestä. Tutkimuksen aineisto koostuu Parsebankista löytyvistä virkkeistä, joissa esiintyy joko sana melkein tai lähes adjektiivin yhteydessä. Tutkimusta varten on saatu valmiiksi seulottu ja annotoitu aineisto (Laippala 2022).

Parsebankista saatava data on annotoitu syntaksijäsentimellä Universal dependencies -skeeman mukaisesti (TNPP). Universal dependencies on yksinkertainen ja yleinen tapa merkitä sanojen välisiä suhteita lauseissa, ja se tarjoaa yhtenäisen tavan kuvailla kielen rakennetta riippumatta kielestä (de Marneffe & al. 2014). Aineisto voidaan analysoida tehokkaasti ja näin eritellä, miten sanat ja lauseen osat liittyvät toisiinsa kieliopillisesti. Tämä mahdollistaa suuren aineiston käyttämisen, kun kaikkea tekstuaalista dataa ei tarvitse annotoida käsin (Luotolahti & al. 2015; Laippala & al. 2018). Profiilien määrittelemiseen hyödynnetään SVM-koneoppimismallilla toteutettua luokittelijaa, joka sekä paljastaa, onko sanojen välillä mahdollisesti eroja, että paljastaa, minkälaisista piirteistä erot koostuvat. Tämä mukailee Laippala et al. (2018) artikkelissaan ehdottamaa koneoppimisen käyttötapaa kielitieteellisen tarkastelun välineenä.

Luokittelijalle on määritelty kaksi luokkaa, *melkein* tai *lähes*, ja se etsii niitä syntaktisia riippuvuussuhdepiirteitä, jotka erottavat sanoja toisistaan. Tämän avulla luokittelija oppii tunnistamaan, kumpi sana tulee kysymykseen missäkin tapauksessa. Luokittelija koulutetaan deleksikalisoidulla eli sanastosta ja sanaluokkainformaatiosta riisutulla biarc-datalla, ja se tekee päätöksensä vain riippuvuussuhteisiin perustuen, ilman sanastollista yhteyttä. Leksikaalisesta yhteydestä on haluttu päästä eroon, jotta syntaktiset piirteet nousisivat pääasiallisiksi. (vrt. Laippala & al. 2018) Erottavia piirteitä voidaan luokittelijan avulla tarkastella ja siten tutkia, millaiset riippuvuussuhteet määrittävät sanojen eroa - millaisia lauserakenteita niihin liittyy.

2 Tutkimuskysymykset

Tässä tutkielmassa lähdetään siitä olettamuksesta, että *melkein* ja *lähes* -sanojen käytön välillä on eroja, jotka kuitenkin ovat niin vaikeasti havaittavia, että kielenpuhujan voi olla vaikeaa eritellä niitä (Edmonds & Hirst, 2008), mutta jotka riippuvuussuhteita kuvastavien biarcien ja

koneoppimiseen perustuvat luokittelijan avulla voidaan saada esiin. Oletuksena siis on, että sanan riippuvuussuhdeprofiilin perusteella voidaan saada tietoa siitä, miksi käyttöön valikoituu tietty sana (ilmaisu).

Tutkimuskysymykset ovat:

1. Vahvistaako SVM:n ja biarcien avulla toteutettu riippuvuussuhdeprofiilianalyysi lähtökohtana olevan hypoteesin, että sanat *melkein* ja *lähes* eivät ole täysin synonyymisia?

2. Mitkä ovat ne luokittelijan avulla löytyvät yleisimmät piirteet ja yhteydet, joiden perusteella käyttöön valikoituu *melkein* tai *lähes*?

3 Aineisto

Tämän tutkimuksen kohteena olevien sanojen *melkein* ja *lähes*, kahden melkein-synonyymisen astemääritteen synonymisuutta tarkastellaan käyttäen aineistona Finnish Internet Parsebankin niitä virkkeitä, joissa jompikumpi sana esiintyy adjektiivin yhteydessä. Virkkeitä aineistosta saatiin yhteensä 16 833, joista 8950 sisälsi sanan melkein ja 7883 sanan lähes. Kielellisesti lauseet edustavat monenlaisia tekstityyppejä ja -lajeja, esim. blogitekstejä, keskustelupalstoilta kerättyjä kommentteja sekä suomenkielisten internet-sivujen tekstejä (Luotolahti et al. 2015).

Yksi suuriaineistoisen korpustutkimuksen mahdollistajista on se, että aineisto voidaan seuloa Finnish Internet Parsebankista, joka koostuu noin 3,7 miljardista sanakkeesta, jotka on automaattisesti analysoitu sekä morfologisesti että riippuvuussuhdesyntaktisesti (*dependency syntax*)(FIP, UD, Turku Neural Parser). Tutkimukseen seuloutuvan aineiston annotoiminen käsin olisi erittäin tehotonta. Näin suurta datamäärää voidaan kuitenkin käyttää, koska annotointi on toteutettu automaattisesti, käsin annotoidun datan avulla koulutetulla jäsentimellä. Merkinnät noudattavat Universal Dependencies -skeemaa, joka on universaaliin syntaktiseen annotaatioon tähtäävä malli (de Marneffe & al. 2014; ks. Luotolahti et al. 2015). Finnish Internet Parsebank koostuu sekä Common Crawlin avulla jo aiemmin kerätystä ei-UD-skeeman mukaan annotoidusta datasta että Luotolahti et al. (2015) toteuttaman web crawl -keräilyn tuottamasta datasta (Luotolahti & al. 2015).

Tutkimuksessa tarkastellaan sanojen syntaktista kontekstia, jota voidaan selvittää jäsentimen prosessoimasta tekstistä. Tutkimuksen aineistosta sanastollinen aines on poistettu ja tarkasteltavia piirteitä ovat vain Universal Dependencies -skeemaan perustuvat

riippuvuussuhteet (UD), joista tutkimuksessa muodostuvat kunkin sanan riippuvuussuhdeprofiilit.

4 Aiempaa tutkimusta

Tämä tutkimus sijoittuu kognitiiviseen kielitieteeseen, jossa tutkimus keskittyy kielenkäyttöön nimenomaan käyttäjän ja käyttöympäristön näkökulmasta, ei rajaten kielen osa-alueita toisistaan (esim. morfologiaan ja syntaksiin) tai asettaen kielen irralleen käyttäjän tietoisuudesta. Kognitiivinen kielioppikäsitys ei tee tiukkaa rajausta kielen ja kielen ulkopuolisen maailman välille (Langacker, 2008), joten nimenomaan käyttötilanteista kerättyyn dataan perustuva tutkimus löytää paikkansa luontevasti juuri tältä kentältä ja kognitiivisen kielitieteen teoreettisesta viitekehyksestä.

5 Melkein-synonymiaan liittyvää tutkimusta

Samalle kentälle tämän tutkielman kanssa sijoittuu esimerkiksi Philip Edmondsin ja Graeme Hirstin tutkimus (2002), joka käsittelee melkein-synonymiaa ja sanavalintaa kielenkääntämisessä. Artikkelissa Edmonds ja Hirst esittelevät laskennallisen mallin melkein-synonyymien hienojakoisuuden hahmottamiselle. He esittävät melkein-synonyymit klusterina, jossa keskiössä on sanojen yhteinen ydinmerkitys, joka jakautuu erilaisiin merkityskenttiin erilaisten periferisten konseptien eli erottavien piirteiden kautta tai vaikutuksesta (Edmonds & Hirst 2002). Näin melkein-synonyymien hienovaraisia eroja voidaan mallintaa ja hyödyntää käännöstieteiden alalla käsiteltäessä oikeaa sanavalintaa käännösten yhteydessä.

Divjak ja Gries (2008) käsittelevät melkein-synonymiaa klusterien muodostamisen kautta. Divjakin ja Griesin tutkimuskohteina ovat venäjän kielen verbit, jotka ilmaisevat yrittämistä, ja työkalunaan käyttäytymisprofiili (*behavioral profile*). Esimerkkejä tutkimuksessa oli 1585 lausetta, jotka annotoitiin käsin. Annotoinnissa käytettiin 87 erilaista ominaisuutta, jotka pitivät sisällään kieliopillista ja syntaktista analyysiä sekä semanttisia parafraaseja, jotka liittyvät mm. ihmisyyden perusalueisiin (“*basic domains*”; Langacker 1987). Näin tutkimuskohteille voitiin luoda moniulotteinen kielelliseen käyttäytymiseen pohjautuva profiili, jonka perusteella niitä saatettiin luokitella ja tarkastella. Divjak ja Gries käyttävät analyysissaan apuna Pearsonin residuaalia, jota yleensä käytetään statistisessa analyysissä vinoumien havainnoimiseen – käänteisesti käyttäen tämä kaava paljastaa semantiikan kannalta nimenomaan olennaisia painotuksia (Divjak & Gries 2008). Pearsonin residuaali kuvaa havaittujen ja ennustettujen arvojen eroa lineaarista regressiomallia käytettäessä. Pearsonin residuaali lasketaan jakamalla yksittäisen residuaalin neliö sen odotusarvolla, ja

käänteinen residuaali taas voidaan arvioida poikkeamia mallin sovittamisessa. Näin poikkeamat siis paljastavat eroja, joita synonymisuuden tutkimisen kohteissa on.

Yksi tutkielman lähteistä ja lähtökohdista on Laippala et al. (2018) artikkeli, joka tarjoaa teoreettiseksi työkaluksi riippuvuussuhdeprofiilin (*dependency profile)* käsitteen sekä käyttää aineistonaan suurta korpusta, Finnish Internet Parsebankia, josta myös oma tutkimusaineistoni on peräisin. Laippala at al. (2018) pitävät mahdollisena, että riippuvuussuhteita kuvastaviin biarceihin perustuva tarkastelu voi paljastaa merkityksellisiä yhteisesiintymiä paremmin kuin leksikaaliset lähestymistavat. Riippuvuussuhdeprofiilit ovat sukua käyttäytymisprofiileille, mutta ne eivät vaadi käsin annotointia, sillä syntaktisten riippuvuussuhteiden määrittely voidaan toteuttaa automaattisesti suurellekin määrälle dataa käyttämällä tarkoitukseen suunniteltua syntaksijäsennintä. Laippala et al. tutkivat suomen kielen konnektiiveja (esim. *eli*, *niin*, *ja*, *mutta* jne.), jotka he ensin luokittelevat klustereihin ja sitten luokittelevat näihin klustereihin perusteella SVM-luokittelijaa käyttäen. Tutkimuksen tarkoituksena on selvittää dependenssiprofiilien käyttöä laajamuotoisessa syntaktisessa analyysissä.

Antti Arppe (2008) on väitöstutkimuksessaan tutkinut ajattelemista merkitsevien sanojen synonymiaa korpuspohjaisesti. Hän yhdistää väitöskirjassa monia vaihteluun perustuvia statistisia tutkimusmetodeja. Arppen tutkimus kohdistuu useaan eri lekseemiin eli sanakkeeseen. Hän esittää, että synonyymisten sanojen tutkiminen pareittain saattaa paljastaa eroja, jotka laajemmassa merkitysklusterissa saattavat osoittautua merkityksettömämmiksi kuin silloin, kun samoja sanoja tutkitaan laajemmassa synonyymisten ilmaisujen ryhmässä (Arppe 2008).

Käännössuomen synonymiasta on kirjoittanut Jarmo Jantunen (2004). Jantunen tutkii käännössuomen erityispiirteitä ja synonyymien käyttöä korpuslähtöisesti. Hänen työkalujaan on kontekstuaalinen profiili, joka perustuu sanojen kontekstissa ilmeneviin yleisiin ilmiöihin. Jantunen tutkii synonymiaa astemääritteiden avulla, ja hän nimeääkin astemääritteet kiinnostaviksi tutkimuskohteiksi: ne ovat yleisiä kielenaineksia, ne sisältävät runsaasti synonymiaa eivätkä ne ole yleensä jakautuneet aihepiirin mukaan. Jantusen lähtökohta on kontekstuaalisessa semantiikassa. (Jantunen 2004.)

6 Astemääritteisiin ja adjektiiveihin liittyvää tutkimusta

2020-luvulla suomen kielen astemääritteistä on kirjoittanut Tuomas Huumo (esim. 2022), joka antaa yleiskuvan astemääritteistä kvanttoreiden määritteenä ja kuvaa ilmiön semanttisia ehtoja. Huumo päättelee, että eri astemääritteiden esiintymiseen vaikuttaa pääsanan laatu.

Lisäksi Huumo (2021) on pohtinut astemääritteiden ominaisuuksia. Huumon tutkimuksissa tarkastelu perustuu pääosin äidinkielisen puhujan kielitajun käyttöön, jonka tukena on ollut myös toisten äidinkielisten puhujien intuitio (Huumo 2021). Huumon tutkimuksen myötä syntyy kysymys siitä, voiko astemääritteisiin liittyviä intuitiivisia tai tutkijan kielitajun mukaisia tuloksia toisintaa tai todistaa suureen korpukseen kohdistuvan analyysin avulla.

Adjektiivien rajallisuudesta (*boundedness*) on kirjoittanut Carita Paradis (2001). Adjektiivien rajallisuus liittyy lingvististen yksiköiden skemaattiseen alueeseen, eli siihen, johon sisältyy yksiköiden konseptuaalinen informaatio (Paradis 2001; ks. esim. Cruse & Togia 1996). Paradis (2001) luokittelee adjektiivit rajoittamattomiin (*unbounded)* ja rajoittuneisiin (*bounded*), joista ensimmäiset ovat skalaarisia, eli ne voidaan ajatella adjektiivin ilmaiseman ominaisuuden jatkumona (esim. *pitkä, hyvä, ilkeä*), kun taas jälkimmäiset ovat joko ääripäitä edustavia (*kauhea, jumalainen*) tai raja-adjektiiveja (*elävä, kuollut*) (Paradis 2001). On huomattava, että jotkin adjektiivit eivät ole laisinkaan rajallisia (*jokapäiväinen, kuvallinen, klassinen*), ja ne eivät saakaan määreekseen astemääritteitä. Paradis käyttää luokittelukriteerinään astemääritteiden laatua, ja siten kiinnostavasti lähestyykin tämän tutkimuksen aluetta ikäänkuin toisesta suunnasta: hänelle adjektiiveja luokitellaan astemääritteiden avulla, kun taas tässä tutkimuksessa nimeomaan astemääritteet ovat kohteena. Huumoon ja Paradisiin pohjaten voidaankin tehdä olettama, että tämän tutkimuksen kohteena olevat approksimatiiviset astemääritteet *melkein* ja *lähes* (termistä ks. Huumo 2021; Paradis 2000) liittyvät lähinnä sellaisiin adjektiiveihin, jotka ovat merkitykseltään joko-tai, eli raja-adjektiiveihin.

7 Perusta ja käsitteet

8 Kognitiivinen kielioppikäsitys

Tutkielman teoreettinen perusta on kognitiivisessa kielioppikäsityksessä. Sen mukaan perinteisen kielioppikäsityksen kielen tarkkarajainen jako leksikkoon ja kielioppiin ei ole mielekäs, koska kieli on olemassa nimenomaan käytettynä eikä irrallaan, teoreettisena kokoelmana eri kategorioita (Langacker 2008). Kognitiivinen kielioppikäsitys on luonnollinen lähtökohta siksi, että tutkimus kohdistuu nimenomaan kielen käyttöön ja perustuu kielen todellisiin käyttöyhteyksiin.

Kognitiivisen kielioppikäsityksen mukaan tiukkaa rajausta kielellisen ja kielen ulkopuolisen tiedon välille ei ole syytä vetää, sillä se olisi väistämättä keinotekoinen. Luonnollinen kieli ei ole itsenäinen ja tarkkarajainen muodollinen systeemi (Langacker 2008). Tämän perusteella ilmaisujen merkitykset määrittyvät kielen esiintymien välisten riippuvuussuhteiden kautta. Kognitiivinen kielioppi on luonnollinen, koska sen keskiössä on merkitys. Toteuttaakseen semiologista funktiotaan – merkityksen symbolistamista fonologisesti – kielessä on oltava kolmenlaisia rakenteita: semanttisia, fonologisia ja symbolisia. Muita rakenteita ei tarvita. Sanasto, morfologia ja syntaksi muodostavat symbolisiin rakenteisiin pelkistyvän jatkumon. (Langacker 2008)

Kognitiivinen kielioppi asettaa sanojen merkityksen kielenkäyttäjään itseensä, eli tämän kognitioon, ja toisaalta esittää ne maailmassa olevien ilmiöiden konseptualisaatioina (Langacker 2008). Merkitys ei siis ole vain kielenkäyttäjän pään sisällä, mutta merkitykset eivät myöskään leijaile käyttäjästä irrallaan jonkinlaisessa kielen sisäisessä tietoisuudessa, vaan merkitys on kielenkäyttäjän käsitteellistämä abstraktio.

Merkityksen konseptuaalisuus antaa pohjan tässä työssä esiteltävälle metodiikalle ja tutkimusongelmille. Koska leksikaalisten yksikköjen (tässä sanojen) merkitys ei ole irrallaan niiden esiintymisyhteyden syntaktisesta todellisuudesta eikä myöskään niiden esiintymisympäristöstä (Langacker 2008), voidaan sanojen merkityseroja, eli vaikkapa niiden keskinäistä synonymisuutta, tarkastella käyttöyhteyksien kautta aineistona todelliset kielenkäyttötilanteet.

9 Synonymia ja melkein-synonymia

Sanojen synonymisuus tarkoittaa niiden läheisintä semanttista samankaltaisuutta (Miller & Charles 1991). Synonymia on perustavanlaatuinen lingvistinen ilmiö, sillä synonymia vaikuttaa leksikon rakenteeseen, samoin kuin melkein-synonymia (Edmonds & Hirst, 2002).

Samanmerkityksisillä sanoilla voidaan sanoa olevan keskeisiä piirteitä, jotka yhdistävät niitä, ja toisaalta perifeerisiä piirteitä, jotka aiheuttavat niiden merkityksen eroavaisuutta. Esimerkiksi sanojen *jänis* ja *pupu* konnotatiiviset eli mielikuviin yhdistyvät merkitykset ovat juuri perifeerisesti eroavia. (Jantunen 2004.)

Absoluuttista synonymiaa voidaan pitää hyvin harvinaisena, ellei jopa mahdottomana ilmiönä (Cruse 1986; Edmonds & Hirst 2002) Synonyymisuuden katsotaan siis joko olevan kontekstiriippuvaista, eli samanmerkityksisten synonyymisten sanojen jakautumista esimerkiksi käyttö- tai asiayhteyksittäin, tai melkein-synonymiaa. Usein edes syntyperäisen kielenpuhujan kyky ei riitä melkein-synonyymien erojen havaitsemiseen ja ilmituomiseen (Edmonds ja Hirst 2002).

Bolingerin (1968) mukaan syntaktisen muodon eroavaisuus heijastaa myös sanojen merkityseroja. Olettamus on, että mikäli sanat eivät ole absoluuttisesti synonyymejä, on niissä oltava jotakin eroa (Edmonds & Hirst, 2002, 108). Lähes samaa tarkoittavien sanojen voi ajatella sijaitsevan lähellä toisiaan, mikä on pohjana myös klustereihin perustuvassa sanaston jaottelussa (esim. Laippala 2018).

Melkein-synonymiaa ei tule ymmärtää vain kahden sanan väliseksi vertailuksi, koska se irrottaa sanat käyttöyhteydestään ja antaa puutteellisen kuvan kielen sanaston muodosta – yhteen merkityskenttään tai merkityksen konseptualisaatioon voi sisältyä monilukuinen määrä toisiaan eri tavoin rajoittavia ja erottavia samaan joukkoon kuuluvia sanoja. Melkein-synonymia on kompleksinen ilmiö, joka vaikuttaa synonymian ja polysemian ohella leksikaaliseen ymmärrykseen. (Edmonds & Hirst 2002, Divjak & Gries 2008)

10 Astemääritteet

Astemääritteisiin viitataan perinteisessä kieliopissa adverbeina (VISK §615). Astemääritteen tehtävä on ilmaista pääsanansa intensiteettiä, joka voi olla matalampi tai korkeampi, ja täsmentää sen skalaarisuuden astetta (Huumo 2022, 87).
Kennedy ja McNally (2005, 367) luokittelevat englannin kieltä käsittelevässä tutkimuksessaan astemääritteet kahteen pääryhmään, avoimen ja sulkeisen skaalan astemääritteisiin, ja Huumo (2022, 95) tämän jaottelun hyväksyen edelleen kategorisoi tutkimukseni kohteena olevan melkein-synonyymiparin *melkein* ja *lähes* approksimatiivisiin astemääritteisiin, jotka ilmaisevat määritettävänsä jäävän niukasti saavuttamatta (*melkein kolmivuotias lapsi, lähes puolityhjä tuoppi*). Sulkeisen skaalan, tai aiemmassa luvussa totaalisuutta määrittäviksi nimetyt astemääritteet määrittävät sellaisia adjektiiveja ”joiden

ilmaisemalla ominaisuudella on minimi- tai maksimiraja tai molemmat” (Huumo 2021; Paradis 2001).

Paradis (2001) jakaa myös omassa tutkimuksessaan astemääritteet kahteen ryhmään perustuen adjektiivien laatuun. Skalaariset astemääritteet (*scalar modifiers*) määrittävät rajoittamattomia adjektiiveja, kuten hyvä tai hieno, kun taas totaalisuutta eli yhtenäistä, rajoittunutta kokonaisuutta ilmaisevien adjektiivien, kuten kanssa esiintyvät totaalisuusastemääritteet (*totality modifiers*). (Paradis 2001.)

11 Käyttäytymisprofiili ja riippuvuussuhdeprofiili

Tutkielmassani esiintyvät käsitteet käyttäytymisprofiili (*behavioral profile*) ja riippuvuussuhdeprofiili (*dependency profile*). Vaikka käsitteet saattavat muistuttaa toisiaan, on kuitenkin syytä ymmärtää niiden eroavaisuus. Käyttäytymisprofiili on useampaan eri piirteeseen ja kokonaisvaltaiseen kontekstiin perustuva psykologiasta lainattu käsite, kun taas riippuvuussuhdeprofiili keskittyy nimenomaan tutkimuskohteen syntaktisten riippuvuussuhteiden luomaan kuvaan.

Kun ajatellaan sanojen merkityksen perustuvan niiden käyttöyhteyksiin, Divjak & Griesin (2008) mukaisesti on käyttäytymisprofiili hyödyllinen käsite. Se muodostuu kohteen piirteistä – millaisten sanojen, rakenteiden jne. kanssa kohdesana esiintyy (Divjak & Gries, 2008). Divjakin ja Griesin (2008) malli perustuu käsin annotoituun dataan, johon on merkitty tietoa sanojen merkityksestä ja semanttisesta riippuvuudesta. Käyttäytymisprofiili pitää sisällään sekä semanttista että morfosyntaktista informaatiota. Tällainen käsitteellistäminen perustuu sanojen (tai ilmaisujen) merkityksen syntymiseen niiden käyttöyhteyksissä ja sopii siksi sanojen merkitysten vertailuun aineistolähtöisesti, käyttäen nimenomaan kielen esiintymiä teoreettisten esimerkkien sijaan. (Divjak & Gries 2008.)

Riippuvuussuhdeprofiilit (*dependency profile, DP*) muodostuvat kohdesanojen yhteisesiintymisistä dependenssisyntaksianalyysistä saadun tiedon kanssa. Tällaisessa syntaktisessa analyysissa on poistettu kaikki muu tieto paitsi syntaktiset riippuvuussuhteet. Riippuvuussuhdeprofiilit eivät vaadi suuria teoreettisia pohjaolettamia, vaan ne perustuvat sanojen yhteisesiintymiin tiettyjen syntaktisten riippuvuussuhteiden kanssa (Laippala et. al. 2018). Sanalla siis ajatellaan olevan tietty syntaktinen ympäristö, jossa se esiintyy. Tämä ympäristö muodostaa sanan riippuvuussuhdeprofiilin. Riippuvuussuhdeteorian muotoilema näkemys on myös, että asioiden (esim. sanojen) esiintymät ovat suoraan suhteessa toisiinsa,

eli vaikkapa astemäärite on läheisemmin suhteessa määrittämäänsä adjektiiviin kuin vaikkapa kyseisen kielen kieliopin formalismiin (Hays 1964).

Riippuvuussuhdeprofiilit ovat hyvä työkalu, koska ne eivät vaadi monimutkaista käsin annotointia. Halutun korpuksen annotaatio voidaan suorittaa lauseenjäsentimellä, ja siksi riippuvuussuhteiden tarkastelu on metodina hyvin yleistettävä.

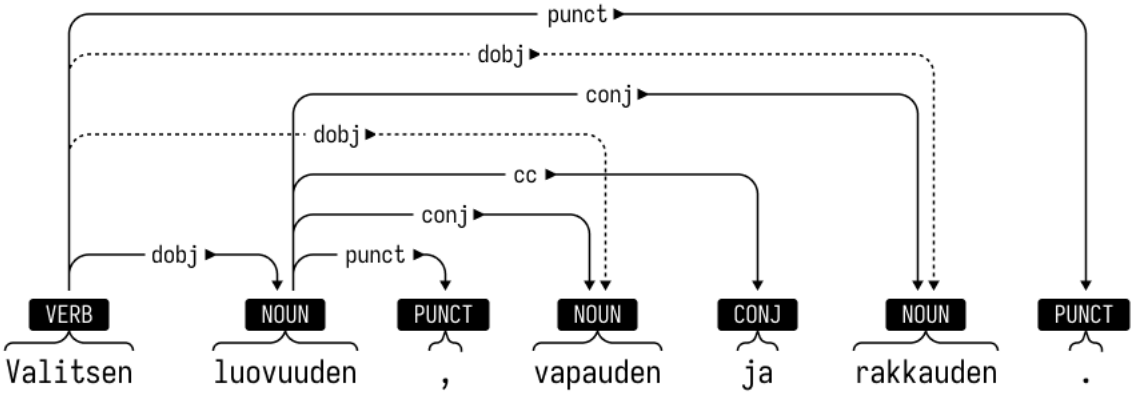
12 Universal dependencies -skeema

Universal Dependencies (UD) -skeema on syntaktisen jäsentämisen standardi, joka tarjoaa yhtenäisen ja formaalin kehyksen eri kielten lauseiden rakenteen kuvaamiseen (ks. de Marneffe & al. 2014). Skeema käyttää morfologisesti ja syntaktisesti merkittyjä riippuvuussuhteita (*dependency relations*). (UD; Luotolahti et al. 2015)

UD-skeeman tarkoitus on luoda yhteinen, koneluettava kieliopillinen kuvaus eri kielten rakenteista, mikä helpottaa kielten vertailua ja kieliteknologisten sovellusten kehittämistä. Se mahdollistaa myös tarkat ja systemaattiset analyysit helpottaen siten esim. poikkikielellisten kielimallien (*cross-lingual models*) kehittämistä. UD-skeemaa kehitetään ja päivitetään jatkuvasti, ja sen kattavuus eri kieliin laajentuu yhä. (UD)

Tutkielman aineiston biarceihin sisältyvä ja niistä saatava informaatio noudattaa UD-skeeman mukaisia merkintöjä, jotka ilmaisevat lauseen sisäisiä syntaktisia riippuvuussuhteita. Tämä on hyödyllinen tapa mallintaa kieltä, sillä se on sekä kielirajoja ylittävää että helposti riisuttavissa sanaluokkainformaatiosta (tai muusta halutusta piirteestä).

Kuvassa 1 jäsenneltynä lause “Valitsen luovuuden, vapauden ja rakkauden.” UD-skeeman mukaisesti. (Luotolahti et al. 2015.)



Kuva 1. UD-skeeman mukaan jäsennetty lause.

13 Aineisto ja menetelmät

14 Tutkimuskohteet *melkein* ja *lähes*

melkein:

vain myönteisissä lauseyhteyksissä: vähää vaille, miltei, lähes, liki, likimain, -määrin, -pitäen, jokseenkin, suunnilleen, osapuilleen, suurin piirtein. (KTS 2021)

lähes:

(kielteisissä yhteyksissä: läheskään) melkein, miltei, liki(pitäen); vrt. likimain, likimäärin, noin, jokseenkin, suunnilleen, osapuilleen, suurin piirtein. (KTS 2021)

Sanalla *melkein* on vastineensa läpi itämerensuomalaisten kielten (SES). Suomen etymologinen sanakirja antaa alkumuodoksi adjektiivin *melkeä* (esiintymä mm. Finno 1580), johon pohjautuva adverbi *melkiäst* löytyy 1500-luvun lopun teksteistä. Agricolalta on löydettävissä *melke(n)* merkityksessä ‘melko, aika’, ja nykyisellään sana esiintyy suomea vastaavassa merkityksessä mm. inkeroisessa melkēn (jossa se saattaa olla suomalainen lainasana), karjalassa *melkie* ‘melko paljon’, *melki* ‘melkein’ ja vatjassa Kukkosin murteessa *melkēttā*, joka voi esiintyä myös kielteisissä yhteyksissä merkityksessä ‘juuri’. On esitetty, että kyseessä olisi muinaisvenäläinen lainasana, jonka nykyedustaja olisi venäjän *mélkij* ’hieno; pieni; matala’.

Lähes taas pohjautuu suomalais-ugrilaiseen sanavartaloon *lähi-*, jonka esiintymiä on laajalti sekä lähi- että etäsukukielissä. Vastineita saattaa olla jopa samojedikielissä asti, jolloin vartalo olisi jo uralilaista perua. Samaan sanueeseen kuuluvat siis *lähellä, lähelle, läheltä.* Vanhastaan *lähes* on voinut merkitä myös ‘lähelle’ (vrt. *ulos, ylös*). (SES)

Kuten edellä olevista kielitoimiston sanakirjan sana-artikkeleista voi päätellä, sanat melkein ja lähes ovat merkityksellisesti erittäin lähellä toisiaan. Ne ovat kumpikin astemääritteitä (Huumo 2021; ISK §615; ks. luku 2.3), tarkemmin luokiteltuna approksimatiivisia astemääritteitä (Huumo, 2021 & 2022). Paradis (2001) jakaa astemääritteet karkeasti kahteen kategoriaan, skalaarisiin (*erittäin, melko*) ja totaalisuuden astemääritteisiin (*täysin, melkein, aivan*), joista jälkimmäiseen tutkimuskohteet kuuluvat. Luokitteluun perustuen myös astemääritteiden pääsanoina esiintyvät adjektiivit ovat merkitykseltään erityyppisiä: approksimatiivisiin astemääritteisiin liittyvät adjektiivit ilmaisevat yleensä raja-arvoa (Paradis 2001; Huumo 2022).

Sanoihin melkein ja lähes liittyy ajatus täyteydestä tai rajasta, jota ei aivan saavuteta. Melkein täysi-ikäinen ei ole vielä aivan saavuttanut 18 vuoden ikää, lähes puolityhjä sali on luultavasti

enemmän kuin puolillaan, jos rajaa lähestytään salin täyteyden näkökulmasta (lähes puolityhjä sali lähestyttynä tyhjyyden näkökulmasta olisi luultavasti puhujan kielitajun vastainen, mikäli sitä ei tulkita jonkinlaisen kielellisen hämäyksen, esimerkiksi huumorin lävitse) (Huumo 2021; Langacker 2008).

15 Aineiston alkuperä: Finnish Internet Parsebank

Finnish Internet Parsebank koostuu noin 3,7 miljardista sanakkeesta, jotka on automaattisesti analysoitu sekä morfologisesti että riippuvuussuhdesyntaktisesti (dependency syntax). Näin suurta datamäärää on mahdollista käyttää, koska annotointi on toteutettu automaattisesti, käsin annotoidun datan avulla koulutetulla parserilla. Merkinnät noudattavat Universal Dependencies -skeemaa, joka on universaaliin syntaktiseen annotaatioon tähtäävä malli. (ks. Luotolahti et al. 2015).

Aineistoa varten korpuksesta seulottiin lauseet, joissa melkein tai lähes esiintyvät adjektiivin astemääreinä. Aineistoon sisältyy vain lauseita, joissa melkein ja lähes esiintyvät juuri edellä mainituissa muodoissa. On mahdollista, että FIP tarjoaisi myös aineistoa, jossa sanojen variantteja (esim. murteellisia) esiintyisi, kuten esimerkiksi murteelliset melkeen tai melkkeist. Lisäksi kyseiseen merkitysklusteriin voi Kielitoimiston sanakirjan perusteella ajatella kuuluvaksi myös esimerkiksi approksimatiiviset astemääritteet likimain, tuskin, liki jne. mutta tutkielman aiheen rajaamiseksi tutkimuskohteiksi otettiin tässä yhteydessä vain yleiskieliset melkein ja lähes.

16 Annotoinnista riippuvuussuhdeprofiileihin

Tutkimusta varten aineiston lauseista on annotoinnin jälkeen muodostettu syntaktiset biarcit (“kaksoiskaaret”; *unlexicalized syntactic biarc*), jotka ilmaisevat aina kahta syntaktista riippuvuussuhdetta, ja kunkin lauseen osalta nämä tallennettiin conllu-muotoon. Lopullisesta aineistosta leksikaalinen data poistettiin, jolloin jäljelle jäävät vain riippuvuussuhteet: näistä biarceista voidaan muodostaa kohdesanan riippuvuussuhdeprofiili perustuen tutkittavan sanan (*melkein* tai *lähes*) ja biarcien väliseen yhteisesiintyvyyteen. Tämä heijastaa esiintymisen lingvististä motivaatiota (Laippala et al. 2018). Laippala at al. (2018) pitävät mahdollisena, että biarceihin eli syntaktisten riippuvuussuhteiden kuvaajiin perustuva tarkastelu voi paljastaa merkityksellisiä yhteisesiintymiä paremmin kuin leksikaaliset lähestymistavat.

Aineistossa lauseita oli kaiken kaikkiaan 16 833, joista 8950 sisälsi sanan melkein ja 7883 sanan lähes, ja biarceja aineistossa oli 630 197, joista erilaisia yhteensä 19 296.

Kuvassa 2 esimerkkikatkelma avatusta CoNLL-U-tiedostosta, jossa näkyvät biarcin lähtösana (1. sarake), biarc (2. sarake; tässä kohtaa sanaluokkainformaatiota ei ole vielä poistettu) ja koko lause tokenisoituna (3. sarake). Kuvassa esillä lause “Sotahan tästä tulee aikanaan jossain syttymään sen voi melkein jo varmaksi sanoa.”

```
tulee   VERB/ROOT/0 ADJ/xcomp/3 VERB/conj/1   sotahan tästä tulee aikanaan jossain syttymään sen voi melkein jo varmaksi sanoa .
tulee   VERB/ROOT/0 PRON/obl/3 VERB/xcomp/1   sotahan tästä tulee aikanaan jossain syttymään sen voi melkein jo varmaksi sanoa .
tulee   PRON/obl/2 VERB/ROOT/0 NOUN/obl/2     sotahan tästä tulee aikanaan jossain syttymään sen voi melkein jo varmaksi sanoa .
tulee   VERB/ROOT/0 AUX/aux/3 VERB/conj/1     sotahan tästä tulee aikanaan jossain syttymään sen voi melkein jo varmaksi sanoa .
tulee   PRON/obl/2 VERB/ROOT/0 VERB/xcomp/2   sotahan tästä tulee aikanaan jossain syttymään sen voi melkein jo varmaksi sanoa .
sanoa   ADV/advmod/2 ADJ/xcomp/3 VERB/conj/0  sotahan tästä tulee aikanaan jossain syttymään sen voi melkein jo varmaksi sanoa .
tulee   VERB/ROOT/0 VERB/xcomp/1 VERB/conj/1  sotahan tästä tulee aikanaan jossain syttymään sen voi melkein jo varmaksi sanoa .
tulee   VERB/ROOT/0 NOUN/obl/1 VERB/conj/1    sotahan tästä tulee aikanaan jossain syttymään sen voi melkein jo varmaksi sanoa .
tulee   PRON/obl/2 VERB/ROOT/0 VERB/conj/2    sotahan tästä tulee aikanaan jossain syttymään sen voi melkein jo varmaksi sanoa .
sanoa   PRON/obj/3 AUX/aux/3 VERB/conj/0      sotahan tästä tulee aikanaan jossain syttymään sen voi melkein jo varmaksi sanoa .
…       …                                     …
```

Kuva 2. Aineiston tiedosto avattuna.

17 SVM-luokittelija

Tutkimuksen aineiston analysoimiseen käytetään ohjattua koneoppimismenetelmää, lineaarista tukivektoriluokittelua (*Linear Support Vector Classification*), ja työkaluna tukivektorikonemalliin (*Support Vector Machine, SVM*) perustuvaa luokittelijaa. Ohjatussa koneoppimisessa mallille syötetään opetusdataa, johon opittava asia on merkitty. Tämän tutkimuksen tapauksessa data siis sisältää luokan (*melkein/lähes*), joka saa arvokseen lauseen syntaktisia riippuvuussuhteita kuvastavat biarcit. Kouluttamisen jälkeen mallin pitäisi pystyä ennustamaan pelkästä biarc-datasta, käytetäänkö lauseessa sanaa melkein vai lähes. Täten siis luokittelijan vastemuuttuja (*response variable*) on lähes/melkein ja ennustemuuttuja (*predictor*) biarcit.

Luokittelija on toteutettu pythonin scikit.learn-kirjastoon sisältyvän LinearSVC-luokan avulla. LinearSVC hyödyntää C/C++:n liblinear-kirjastoa, joka on suunniteltu suurten lineaaristen oppimistehtävien ratkaisemiseen ja se keskittyy erityisesti lineaarisiin tukivektorikoneisiin ja logistiseen regressioon. Yksi sen tärkeimmistä eduista on sen laskentatehokkuus suurten datakokonaisuuksien (*dataset*) käsittelyssä, mikä tekee siitä erityisen sopivan tilanteisiin, joissa esimerkkidataa tai piirteitä on paljon. LinearSVC-luokan avulla toteutettu tukivektorikone on ytimeltään (*kernel*) lineaarinen, mikä tarkoittaa, että luokittelija etsii luokkien välille suoraviivaista jakolinjaa. (Pedregosa et al. 2011)

18 Luokittelijan arviointi

Tukivektorikoneen arvioimisessa käytetään yleisimmin neljää mittaria: osuvuus (accuracy), tarkkuus (*precision*), herkkyys (*recall*) ja F1-pisteytys (*F1-score*). Lisäksi hyödyllinen mittari

on sekaannusmatriisi (*confusion matrix*). Tässä työssä esittelen tuloksia näihin mittareihin pohjaten.

Osuvuus kuvaa sitä, montako kertaa kone ennustaa luokan oikein, eli se on oikeiden ennustettujen suhde kaikkiin ennustuksiin. Jos esimerkiksi ennustuksia on 100, ja luokittelija ennustaa 40 oikein yhteen luokkaan ja 40 toiseen, olisi mallin osuvuus tuolloin 0,8. Binäärisen luokittelijan, eli kaksi luokkaa sisältävän luokittelijan, perustaso on 0,5, kun aineistossa kutakin luokkaa edustaa sama määrä esimerkkejä. Tässä tutkimuksessa luokat ovat 0=lähes ja 1=melkein, ja kouluttamiseen käytetään tasapainotettua aineistoa, eli kumpaakin luokkaa edustaa yhtä monta esimerkkiä.

Tarkkuus kuvastaa oikein ennustettujen määrää suhteessa kaikkiin luokkaan kuuluviksi ennustettujen määrään, ja herkkyys taas lasketaan vertaamalla oikein ennustettujen määrää kaikkiin aineistossa oleviin oikeisiin vaihtoehtoihin. F1 on edellä esitettyjen mittojen harmoninen keskiarvo.

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\,\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{tp}}{2\mathrm{tp} + \mathrm{fp} + \mathrm{fn}}$$

Kuva 3. F1-arvon laskeminen. Kuvan lähde: Wikipedia.

Aineisto jaettiin satunnaisesti opetusjoukkoon (*training set*) (80 %) ja testausjoukkoon (*test set*) (20 %). Mallin sisäinen tarkkuus (*precision*) ja herkkyys (*recall*) laskettiin testausjoukon avulla. Tämä toistetaan riittävän monta kertaa, jotta voidaan varmistaa, ettei aineisto ole sattumalta jakautunut epätasaisesti opetus- ja testausjoukon kesken (Laippala et al. 2018). Tätä työtä varten malli ajettiin sata (100) kertaa.

19 Prosessi

20 Datan valmistelu

Ohjattu koneoppiminen vaatii myös koulutukseen käytettävältä datalta selkeää muotoa. Luokittelijan kouluttamista varten dataa käsiteltiin, jotta luokittelijan olisi helppo käyttää sitä. Tutkimusta varten Finnish Internet Parsebankista seulotut, lähes- tai melkein-sanoja adjektiivin astemääritteenä sisältävät lauseet ja niistä muodostetut biarcit saatiin jäsenneltyinä CoNLL-U.gz-tiedostoihin.

Koulutus- ja testidatan käyttämistä varten tiedostoista kerättiin vain biarcit python-sanakirjoihin, jotka koottiin listaksi. Sanakirjoissa avain (*key*) määrittää luokkaa, ja sen arvoina (*value*) ovat biarcit.

Tämän jälkeen biarceista poistettiin leksikaalinen informaatio, eli sanaluokkia osoittavat osat. Tämä tehtiin, jotta luokittelija voitaisiin kouluttaa nimenomaan puhtaasti riippuvuussuhteisiin perustuvalla datalla. Lisäksi biarcit tiivistettiin yhdeksi piirteeksi poistamalla niistä välilyönnit, jotta välilyöntiin perustuva tokenisoija (*tokenizer*) pitäisi biarcit kokonaisina, ja luokat muutettiin muotoon ‘0’ ja ‘1’. Luokka 0 osoittaa arvon kuuluvan luokkaan *lähes*, 1 puolestaan luokkaan *melkein*. Kuvat 3, 4, 5 ja 6 osoittavat datan käsittelyprosessia; kuvissa samaa lausetta kuvavaat biarcit. Kuvissa esillä lause “Sotahan tästä tulee aikanaan jossain syttymään sen voi melkein jo varmaksi sanoa.”

```
data = [
    {
        'melkein' : [
            'VERB/ROOT/0 ADJ/xcomp/3 VERB/conj/1',
            'VERB/ROOT/0 PRON/obl/3 VERB/xcomp/1',
            'PRON/obl/2 VERB/ROOT/0 NOUN/obl/2',
            'VERB/ROOT/0 AUX/aux/3 VERB/conj/1',
            'PRON/obl/2 VERB/ROOT/0 VERB/xcomp/2',
            'ADV/advmod/2 ADJ/xcomp/3 VERB/conj/0',
            'VERB/ROOT/0 VERB/xcomp/1 VERB/conj/1',
            'VERB/ROOT/0 NOUN/obl/1 VERB/conj/1',
            'PRON/obl/2 VERB/ROOT/0 VERB/conj/2',
            'PRON/obj/3 AUX/aux/3 VERB/conj/0',
            ...
        ]
    },
    ...
]
```

Kuva 3. Vaihe 1. Biarc-data tuotuna python-sanakirjaan.

```
data = [
    {
        'melkein' : [
            'ROOT/0 xcomp/3 conj/1',
            'ROOT/0 obl/3 xcomp/1',
            'obl/2 ROOT/0 obl/2',
            'ROOT/0 aux/3 conj/1',
            'obl/2 ROOT/0 xcomp/2',
            'advmod/2 xcomp/3 conj/0',
            'ROOT/0 xcomp/1 conj/1',
            'ROOT/0 obl/1 conj/1',
            'obl/2 ROOT/0 conj/2',
            'obj/3 aux/3 conj/0',
            ...
        ]
    },
    ...
]
```

Kuva 4. Vaihe 2. Biarc-data riisuttuna sanaluokkainformaatiosta.

```
data = [
    {
        'melkein' : [
            'ROOT/0xcomp/3conj/1',
            'ROOT/0obl/3xcomp/1',
            'obl/2ROOT/0obl/2',
            'ROOT/0aux/3conj/1',
            'obl/2ROOT/0xcomp/2',
            'advmod/2xcomp/3conj/0',
            'ROOT/0xcomp/1conj/1',
            'ROOT/0obl/1conj/1',
            'obl/2ROOT/0conj/2',
            'obj/3aux/3conj/0',
            ...
        ]
    },
    ...
]
```

Kuva 5. Vaihe 3. Biarc-data, josta välilyönnit on poistettu.

```
data = [
    {
        '1': 'ROOT/0xcomp/3conj/1 ROOT/0obl/3xcomp/1 obl/2ROOT/0obl/2 ROOT/0aux/3conj/1 obl/2ROOT/0xcomp/2 ...'
    },
    ...
]
```

Kuva 6. Vaihe 4. Biarc-data valmiina välilyönnillä erottelevaa tokenisoijaa varten.

SVM:n kouluttamista varten valittiin kummastakin luokasta satunnaisesti 7500 esimerkkiä, jotta data olisi tasapainoinen, eli kumpaakin luokkaa edustaisi sama määrä instansseja. Nämä esimerkit sisältävät 561 385 biarcia, eli noin 37 biarcia/lause. Yksittäisiä biarceja aineistossa oli 18459.

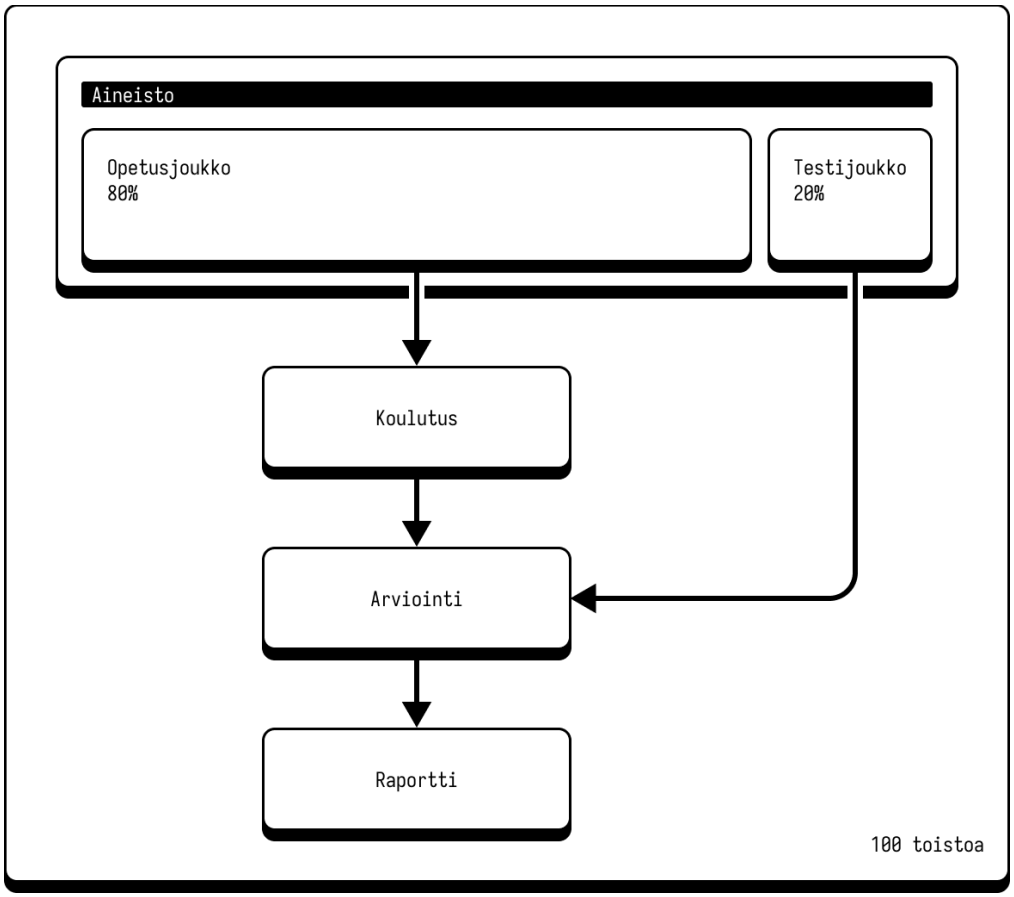
21 Hyperparametrit

Tutkimusta varten luotiin lineaarinen tukivektorikone, jonka rakennetta on kuvattu yllä. Sen hyperparametreista *C-arvoksi* valikoitui 0.05, jolla luokittelijan tarkkuus oli parhaimmillaan. Lisäksi hyperparametriä *tol* (=toleranssi) säädettiin arvoon 1e-5 (eli 0.00001:een). *Tol*-hyperparametri määrittää sitä toleranssia tai "sietokykyä", jonka avulla malli arvioi, koska optimointi-iteraatiot eli kouluttamisen voi lopettaa. Toleranssin pienentäminen lisää vaatimuksia mallin suorittamiseen vaadittavalle suoritusteholle, mutta se tuottaa yleisesti ottaen tarkemman ratkaisun. Tämän tutkimuksen yhteydessä tehovaatimus ei ollut kynnyskysymyksenä, mutta se on syytä ottaa huomioon, jos käsitellään vielä huomattavasti suurempaa aineistoa.

Muutoin luokkaa LinearSVC edustavan tukivektorikoneen hyperparametrit ovat luokan oletusarvojen mukaisia. Hyperparametrit ovat siis (muutetut **lihavoitu**):

class sklearn.svm.LinearSVC*(penalty='l2', loss='squared_hinge', *, dual='warn',* ***tol=0.00001***, ***C=0.5***, *multi_class='ovr', fit_intercept=True, intercept_scaling=1, class_weight=None, verbose=0, random_state=None, max_iter=1000)*

22 Juoksutus



Kuva 7. Juoksutuskaavio

Aineisto jaettiin satunnaisesti opetusjoukkoon (*training set*) (80 %) ja testausjoukkoon (*test set*) (20 %). Mallin sisäinen tarkkuus (*precision*), herkkyys (*recall*) ja F1-arvo (*F1-score*) laskettiin testausjoukon avulla. Lisäksi laskettiin vielä sekaannusmatriisi (*confusion matrix*).

Mallin koulutus ja arviointi toistettiin yhteensä 100 kertaa. Jokaisen kierroksen yhteydessä data jaettiin uudelleen satunnaisesti opetus- ja testausjoukkoon, jotta voitiin varmistaa, ettei aineisto ole sattumalta jakaantunut epätasaisesti joukkojen kesken. Useampi iteraatio on myös tarpeen, jotta tulokset eivät pohjautuisi mahdolliseen häiriöön, kuten koulutus- ja testausjoukon epätasaiseen jakautumiseen (Laippala et al. 2018).

Jokaisesta iteraatiosta tallennettiin raportti .txt-tiedostoon. Raportit sisältävät tiedon mallin tarkkuudesta kyseisellä ajokerralla, luokitteluanalyysin (sisäinen tarkkuus, herkkyys ja F1)

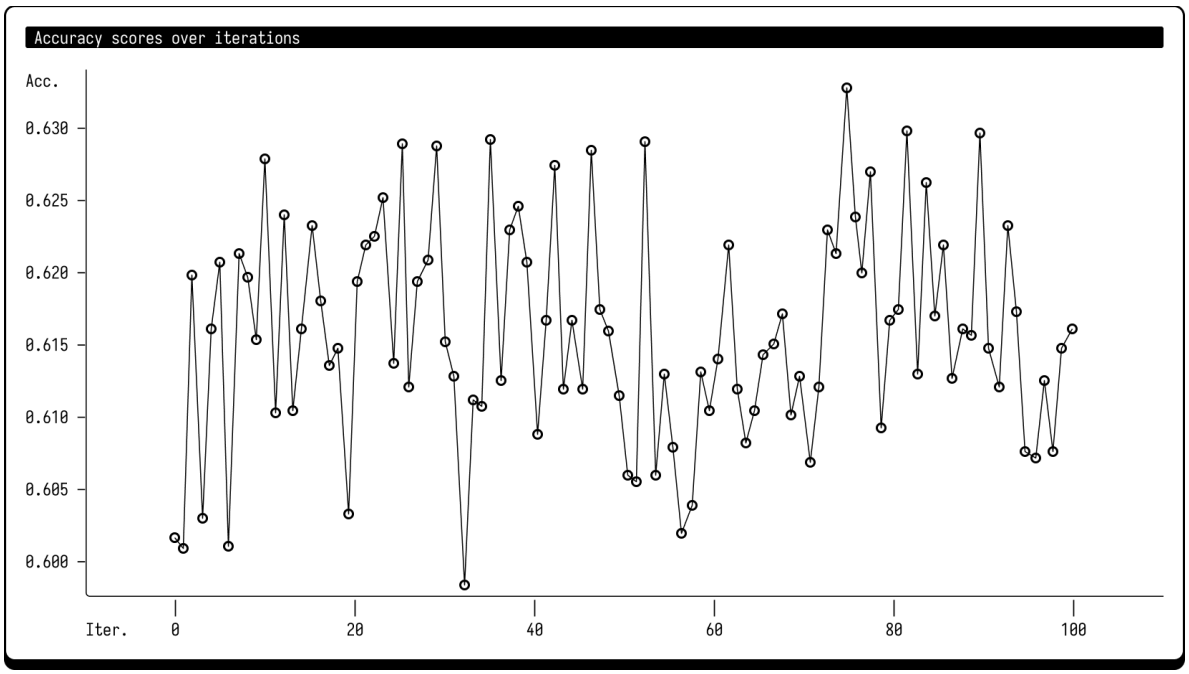
sekä sekaannusmatriisin. Lisäksi jokaiseen raporttiin on kirjattu kunkin ajokerran mukaiset 50 yleisintä piirrettä painokertoimineen (*coefficient*) ja vektoreineen. Lisäksi jokaiselta kerralta seulottiin 20 useimmin esiintyvää biarcia ja näistä määriteltiin kutakin luokkaa yleisimmin kuvaavat biarcit, eli ne, jotka olivat useimmin päässeet 50 yleisimmän erottavan piirteen joukkoon.

23 Tulokset ja analyysi

Tässä luvussa käsittelen ensin luvussa 3.4.1. esitettyjä SVM:n arviointikriteerien mukaisia tunnuslukuja. Ne osoittavat ensimmäiseksi, onko malli löytänyt luokkien välille minkäänlaisia eroja perustuen aineistona annettuihin biarceihin. Sen jälkeen esittelen yksittäisiä luokkia erottavia piirteitä, eli yksittäisiä deleksikalisoituja biarceja, ja lopulta pilkon biarceja osiin selvittääkseni, minkälaisia syntaktisia riippuvuussuhteita kumpikin luokka pitää sisällään perustuen 20 yleisimpään biarciin.

24 Luokittelijan arviointi

Sadan iteraation jälkeen parhaimman suorituksen osuvuus oli 0,633. Heikoimman iteraation suorituksen osuvuus oli puolestaan 0,598. Kuviosta 1 voidaan nähdä suoritusten vaihtelu eri iteraatiokerroilla.



Kuvio 1. Mallin suorituskyvyn vaihtelu iteraatioiden aikana.

Koska kahden luokan luokittelijan perustaso on 0,5, kun aineistoa on kummastakin luokasta yhtä paljon, voidaan todeta mallin huonoimillaankin oppineen jonkin verran tunnistamaan luokkia. Luokkien välillä siis on eroavaisuuksia, joita voidaan havainnoida ainakin tukivektorikoneeseen perustuvan luokittelijan avulla. Tämä vaikuttaisi vahvistavan tutkimuksen lähtökohdan, että melkein ja lähes eivät tosiaan ole täysin synonyymisia, vaan

biarcien avulla tarkasteltujen käyttöyhteyksien perusteella niiden voidaan väittää tosiaan olevan melkein-synonyymeja.

Classification report — Accuracy 0.633

	Precision	Recall	F1-score	Support
0 = lähes	0.63	0.64	0.64	1500
1 = melkein	0.64	0.62	0.63	1500
Avg./total	0.63	0.63	0.63	3000

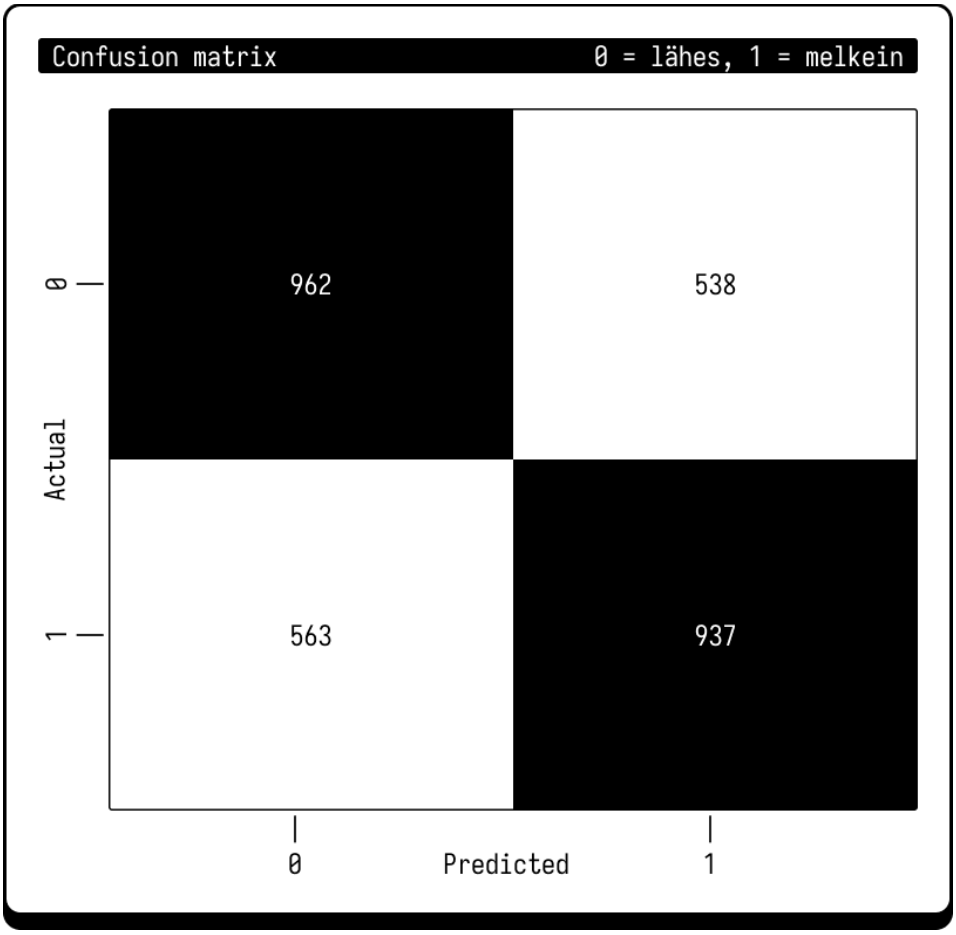
Kuva 8. Parhaimman iteraation tunnusluvut, osuvuus 0,633.

Classification report — Accuracy 0.598

	Precision	Recall	F1-score	Support
0 = lähes	0.60	0.60	0.60	1500
1 = melkein	0.60	0.59	0.60	1500
Avg./total	0.60	0.60	0.60	3000

Kuva 9. Huonoimman iteraation tunnusluvut, osuvuus 0,598.

Kuvissa 8 ja 9 olevissa taulukossa esillä sekä parhaimman että huonoimman iteraation tunnusluvut kahden desimaalin tarkkuudella. Luokittelija näyttää tunnistavan yhtä hyvin kummankin luokan edustajia, ja f1-arvo onkin samaa luokkaa kuin tarkkuus. Parhaassa tuloksessa luokittelijan f1-luku on luokan 0=lähes 0,64 ja luokan 1=melkein 0,63, ja huonoimmassa se asettuu kummallekin luokalle 0,60:een. Merkittävää eroa ei siis luokkien tunnistamisen välille synny. Oikeita osumia malli tunnisti parhaimmillaan luokasta 0=lähes 962, vääriä taas 563; luokasta 1=melkein oikeita osumia tuli 937 ja vääriä 538. Kuvassa 11 sekaannusmatriisi avattuna kaavioksi.



Kuva 11. Sekaannusmatriisi, kun tarkkuus on 0,633 (paras iteraatio)

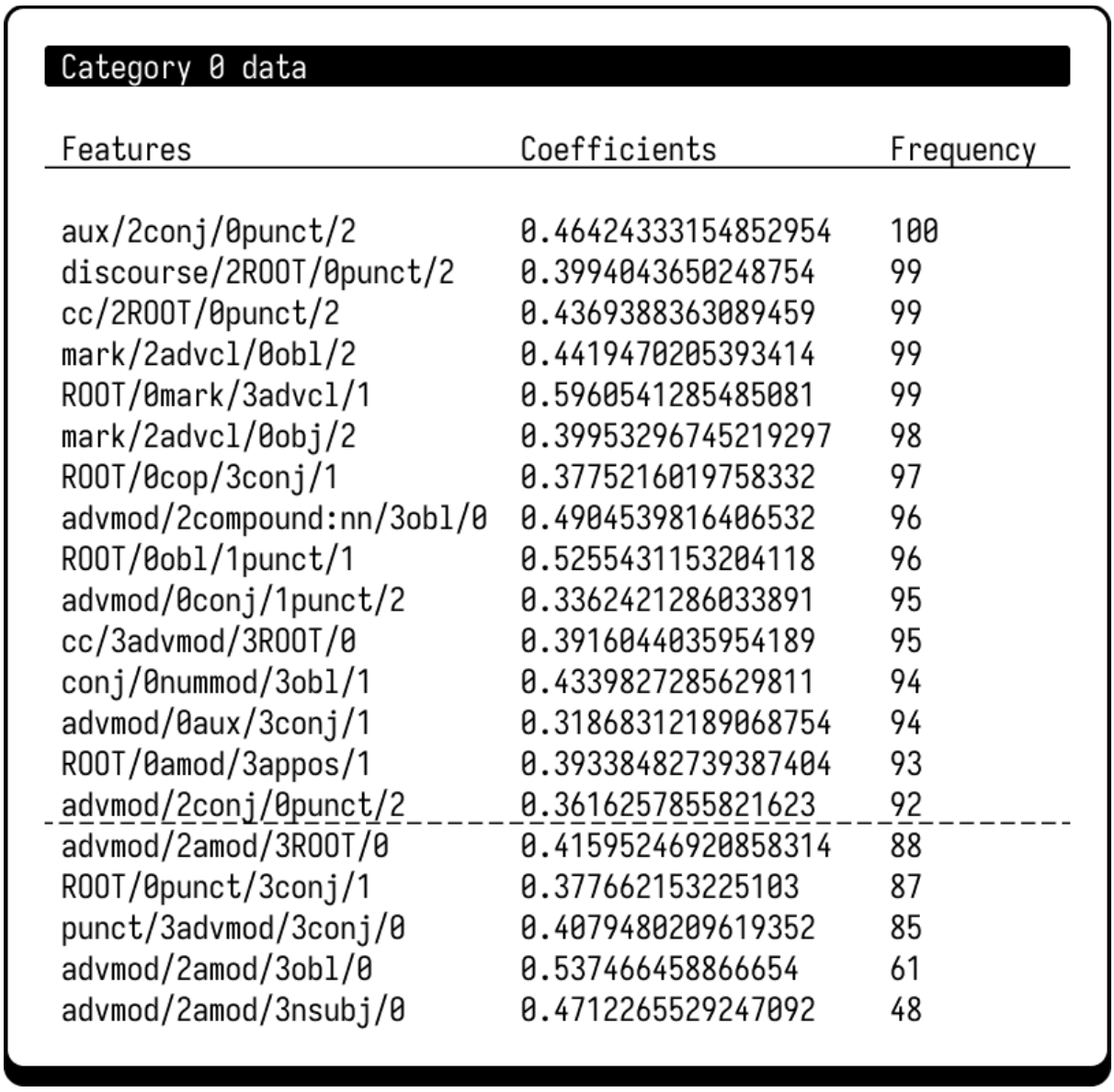
25 Analyysi

Jokaisen iteraation kertoimeltaan korkeimmasta 50 biarcista lasketuista yleisimmästä 20 biarcista on syytä tarkemmin tarkastella sitä, mitkä piirteet ovat yleisimmin toimineet määrittämässä luokkaa iteraatioissa ja mitkä ovat niiden keskimääräiset coef-kertoimet. Alla olevassa taulukoissa kuvissa 12 ja 13 näkyvät myös esiintymisten määrä sadan iteraation aikana. Piirteiden kertoimet (*coefficients*) esittävät painoa, jolla piirre määrittää luokkaa. Positiiviset kertoimet tarkoittavat, että piirre kuvastaa luokkaa 0=*lähes*, kun taas negatiiviset kertoimet osoittavat luokkaa 1=*melkein*. Kaikki ne piirteet, joilla on muu kerroin kuin 0, ovat osana määrittämässä luokittelijan päätöstä siitä, kumpi sana tulee kyseeseen.

Biarcien suhteellista merkittävyyttä voidaan arvioida esimerkiksi sillä, kuinka usein ne esiintyvät luokkaa määrittävinä piirteinä iteraatioiden aikana (Laippala et al. 2018). 20 yleisimmän kunkin luokan piirteistä voidaan huomata, että osa piirteistä on noussut tuohon joukkoon lähes joka kierroksella. Usein esiintyminen ei kuitenkaan välttämättä korreloi kertoimen kanssa. Niinpä kerroin saattaa olla korkea, mutta siitä huolimatta piirre nousee luokan määrittäjäksi harvemmin, kuten esimerkiksi kuvassa 12 esiintyvä biarc

'advmod/2amod/3obl/0', kertoimeltaan 0.537466458866654 on valittu 61 kertaa; piirteen kerroin on ryhmänsä korkeimpia, mutta se on valikoitunut erottamaan luokkaa vain 61 % iteraatioista. Tämä saattaa heijastella sitä, että jotkin biarcit olivat tärkeitä vain osassa aineistoa (Laippala et al. 2018). Otankin tarkasteluun nyt ne piirteet, jotka ovat ohjanneet luokittelijaa yli 90 kertaa sadasta. Tällaisia piirteitä on 15 kummassakin ryhmässä.

Tarkastelen biarceja niiden osasten kautta, ja tarkastelen näitä piirteitä esimerkein kunkin luokan osalta. Kategorioihin ohjaavista piirteistä voidaan etsiä luokan sisäisiä yhtäläisyyksiä tai luokkien välisiä eroja tarkastelemalla niissä usein esiintyviä riippuvuussuhteita ja näiden yhteisesiintymiä. Nostan tässä seuraavissa luvuissa esiin sellaisia piirteitä, jotka vaikuttavat kunkin luokan kohdalla merkittäviltä. Kiinnostavia osia sisältäviä biarceja voidaan tarkastella kokonaisuudessaan, ja niiden avulla voidaan palata aineiston lauseisiin, jotka esikäsittelyssä poistettiin. Esimerkkilauseet esitetään kukin biarcinsa ja biarcin lähtösanan (eli lähtöpisteen, jota merkitsee biarcissa luku 0 piirteen jälkeen, esimerkiksi conj/0) kanssa.



Category 0 data

Features	Coefficients	Frequency
aux/2conj/0punct/2	0.46424333154852954	100
discourse/2ROOT/0punct/2	0.3994043650248754	99
cc/2ROOT/0punct/2	0.4369388363089459	99
mark/2advcl/0obl/2	0.4419470205393414	99
ROOT/0mark/3advcl/1	0.5960541285485081	99
mark/2advcl/0obj/2	0.39953296745219297	98
ROOT/0cop/3conj/1	0.3775216019758332	97
advmod/2compound:nn/3obl/0	0.4904539816406532	96
ROOT/0obl/1punct/1	0.5255431153204118	96
advmod/0conj/1punct/2	0.3362421286033891	95
cc/3advmod/3ROOT/0	0.3916044035954189	95
conj/0nummod/3obl/1	0.4339827285629811	94
advmod/0aux/3conj/1	0.31868312189068754	94
ROOT/0amod/3appos/1	0.39338482739387404	93
advmod/2conj/0punct/2	0.3616257855821623	92
advmod/2amod/3ROOT/0	0.41595246920858314	88
ROOT/0punct/3conj/1	0.377662153225103	87
punct/3advmod/3conj/0	0.4079480209619352	85
advmod/2amod/3obl/0	0.537466458866654	61
advmod/2amod/3nsubj/0	0.4712265529247092	48

Kuva 12. Luokan 0=lähes yleisimmät piirteet (features), kertoimet (coefficients) ja yleisyys (frequency)

26 Lähes

Luokan 0 eli lähes syntaktisista riippuvuussuhteista viisi kertaa tai useammin esiintyvät conj eli konjunkti, advmod eli adverbiaalimäärite, *punct* eli punktuaatio tai välimerkki ja *ROOT* eli lauseen juurisana. Näistä ainoastaan punct on sellainen, jota ei esiinny toisessa luokassa. Tarkastelen esimerkkejä näitä osasia sisältävistä biarceista lähtöisin; lisäksi melkein-luokassa tulevat kyseeseen myös sellaiset syntaktiset suhteet, joita ei esiinny luokan lähes biarceissa.

Voidaan huomata, että conj eli konjunkti esiintyy tarkasteltavissa piirteissä 3 kertaa adverbiaalisen määritteen (*adverbial modifier* = *advmod*) kanssa sekä kerran numeerisenn määritteen (*numeric modifier* = *nummod*) kanssa. Konjunkti tarkoittaa kahden elementin välistä suhdetta, joka muodostuu sanojen välille konjunktion kautta. Konjunktin yleinen

esiintyminen voi kertoa pidemmistä lauseista, sillä esimerkiksi rinnasteisten päälauseiden verbit ovat konjunkteja. Toisaalta konjunktit liittyvät myös listaamiseen, sillä listan ensimmäistä elementtiä seuraavat osat luetaan ensimmäisen konjunkteiksi (UD).

Adverbiaalimäärite on myös yleisempi luokassa *lähes* kuin luokassa *melkein*. Adverbiaalimäärite on adverbi tai adverbiaalilauseke, joka vaikuttaa verbiin tai määritesanoihin, kuten adjektiiveihin tai adverbeihin. Adjektiivin astemääritteenä *lähes* ja *melkein* tulevat määritellyksi nimenomaan adverbiaalimääritteiksi. Tästä johtuen *advmod* esiintyy ainakin kerran joka esimerkkilauseessa. Se voi silti liittyä luokkaa määrittäviin piirrebiarceihin, koska ne sisältävät kolme sanaa ja erilaisia yhdistelmiä.

Riippuvuussuhde *mark* eli alisteista lausetta ilmaiseva sana, kuten alistuskonjunktio, esiintyy vain luokassa lähes. Tämä saattaa osoittaa samaa kuin konjunktin yhteydessä hahmoteltu ajatus mahdollisesta lauseen pituuden vaikutuksesta sanavalintaan *lähes* ja *melkein* välillä.

Lähes-luokan piirteissä on kaksi biarcia, jotka pitävät sisällään saman kokonaisuuden, *advmod*, *conj* ja *punct*, mikä tarkoittaa, että näiden biarcien osana on myös välimerkkejä. Välimerkkeihin liittyviä riippuvuuksia ei esiinny useimmin luokkaa määrittävissä piirteissä lainkaan luokassa *melkein*, joten tämä yhdistelmä on luokkaan lähes liittyvä vahvasti ohjaava piirre. Välimerkkejä ei varsinaisesti UD-skeeman mukaan pidetä tavallisena riippuvuussuhteena, vaan ne ovat aina sidoksissa toiseen leksikaaliseen yksikköön (sanaan), eivätkä ne voi toimia pääsanana (UD). Silti tämän tutkimuksen asetelmassa niiden tarkastelu on sikäli kiehtovaa, että myös välimerkkien määrä voi kertoa jotakin sanojen melkein ja lähes käyttöyhteyksistä kirjallisesti tuotetussa kielessä, vaikkakin välimerkkien käytön voi ajatella liittyvän ei ainoastaan kielenkäyttöön vaan myös kirjoituskontekstiin ja vaikkapa oikeinkirjoitustaitoon.

Esimerkeissä 1 ja 2 nähdään kaksi lausetta, joissa esiintyy lähes-luokalle tyypillinen yhdistelmä, *advmod*, *conj* ja *punct*. Toisen lauseen *conj*-suhdetta edustaa adverbi, toisessa verbi. Esimerkeissä ensin deleksikalisoitu biarc, sitten biarcin lähtösana ja sen jälkeen lause, jossa biarciin liittyvät sanat on lihavoitu.

Esimerkki 1: advmod/2conj/0punct/2 ***ähkytäynnä***

unta riittäis **vaikka** 24h **,** syödä vois lähes koko ajan (tai välillä on yhdestä leivästä ihan **ähkytäynnä**) ja huimaus vaivaa .

Esimerkki 2: advmod/2conj/0punct/2 ***tilasit***

" älä saatana , mitä sä **oikeen tilasit ?** " , chester loi lähes sympaattisen katseen januaryyn joka oli todennut juomansa olevan lähes myrkytetyn makuista .

Esimerkissä 3 taas mukana *nummod* eli numeerinen määrite, joka siis ilmaisee lukumäärää numeroin (UD).

Esimerkki 3: conj/0nummod/3obl/1 ***teen***

(nyt on pikkuinen ihmistyttö ja paljon opiskeluja..rajoittaa kovin) varmaankin kuukauden yhteensä sitä näpertelin , mutta verraten ompeluun , kun pääsee lähes koko päivän touhuamaan , niin **teen** asun ehkä **3-5 päivässä** .

Yllä olevissa esimerkeissä *conj*-riippuvuussuhde on aina biarcin alkuna tai lähtösanana, mutta se esiintyy yhtä hyvin muussakin kohtaa biarcia, esimerkiksi seuraavassa esimerkissä:

Esimerkki 4: advmod/0aux/3conj/1 ***jäljellä***

päätöserän alussa johtolukemat edelleen 1-3. tosin vajaa erä **jäljellä** , mutta jo tähän mennessä otteet **ovat** olleet lähes **sensaatiomaiset** .

Esimerkissä 5 riippuvuussuhde *mark* kohdistuu rinnastuskonjunktioon *mutta.*

Esimerkki 5: mark/2advcl/0obl/2 ***pistän***

sensorointikäyrissä näkyy hyvin ihan huipputasaiset yöt , lähes suoraa viivaa ja verensokerit 6-8 , **mutta** annas olla kun **pistän** ekan leipäpalan suuhun !

27 Melkein

Melkein-luokkaa yleisimmin määrittävien biarcien (kuvassa 13) osasista esiin nousevat piirteiksi nousevat *amod*, *nmod*:*poss* sekä *nsubj*. *Amod*, eli adjektiivinen määrite, joka määrittää substantiivia tai pronominia, muttei lausetta, esiintyy luokan 15 yleisimmästä

piirteestä kuudessa ja yhteensä seitsemän kertaa. Luokassa lähes *amod* esiintyy vain kerran, joten se on selvästi vallitsevampi piirteen osa *melkein*-ryhmässä.

Nmod:poss eli ilmaisee possessiivista nominaalimääritettä, joka on pääsanaansa omistussuhteessa. Se on nominaalimääritteiden alaluokka (tai tarkennus), joka ilmaisee nimenomaan omistusta, ei omistusta ilmaisevaa appositiota. Tätä riippuvuussuhdetta ei ole *lähes*-luokan yleisimmissä piirteissä lainkaan.

Nsubj ei esiinny luokassa *lähes* lainkaan. Se viittaa nominaalisubjektiin, joka liittyy usein suoraan juuriverbiin tai sen konjunkteihin. Nominaaliseen subjektiin liittyvät syntaktiset piirteet näyttävät puolestaan olevan hyvin edustettuna melkein-valintaan ohjaavina piirteinä.

Esimerkissä 6. esillä lause, joka sisältää esitellyistä riippuvuussuhteista nominaalimääritettä *nmod:poss* ja *amod*. Esimerkissä biarcin lähtösananana on riippuvuus obj, joka liittyy objektiin. Possessiivinen nominaalimäärite on esimerkissä erisnimi, eli substantiivi.

Esimerkki 6: nmod:poss/3amod/3obj/0 ***talot***

hilding ekelund suunnitteli **porintie 5:n punatiiliset talot** , joiden keskellä on melkein umpinainen sisäpiha .

Esimerkissä 7 ja 8 puolestaan lauseita, joissa esiintyy *nsubj.* Toisessa sen asemassa on substantiivi, toisessa pronomini.

Esimerkki 7: nsubj/2ROOT/0obj/2 ***nukkuu***

vauva nukkuu pitkät **yöunet** , mutta sitten syö melkein koko ajan päivällä .

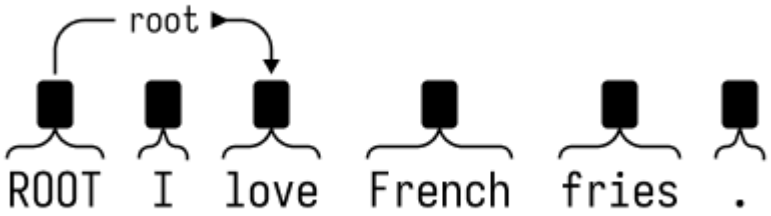
Esimerkki 8: nsubj/2ROOT/0obj/2 ***tuo***

tämä tuo tekstiin paikoin melkein sciencefiktiivistä **ulottuvuutta** ja saa lukijan jännittämään monen kymmenen sivun verran sitä , kuinkahan tässä käy .

28 Yhteisiä, erottavia piirteitä

29 ROOT

Riippuvuussuhde *ROOT* edustaa koko lauseen juurta. Se on kieliopillinen apuväline, joka ei ole varsinaisesti lauseen sana, vaan kuvitteellinen jäsennyksen solmukohta, josta suhde lauseen pääsanaan, useimmiten verbiin, saadaan muodostettua. Kuvassa 14 esitetään ROOTin sijainti suhteessa lauseeseen.



Kuva 14. ROOT kuvattuna lauseessa *I love french fries*, suom. rakastan ranskalaisia perunoita. ROOT liittyy verbiin *love*.

Tästä riippuvuussuhteesta johtuen *ROOT* esiintyy hyvin usein määrittävissä piirteissä (ja aina asemassa 0), mutta huomattavaa on se, minkä kanssa se esiintyy. Kuvassa 15 olevasta taulukosta selviää, miten muut riippuvuussuhteet jakautuvat yhdistettynä *ROOTiin.*

lähes	molemmat	melkein
discourse	amod	nsubj
cc	obl	obj
punct		acl
advcl		nsubj:cop
appos		compound:nn
advmod		
conj		
mark		
cop		

Kuva 15. Riippuvuussuhteiden jakautuminen yhdistettynä ROOTiin.

Yhteisiä *ROOTiin* liittyviä riippuvuussuhteita yleisimmissä piirteissä ovat vain aikaisemmin selvitetty *amod* ja *obl.* Esimerkeissä 9 ja 10 *lähes*- ja *melkein*-luokkien esimerkit lauseista, joissa on niiden ominaisia *ROOTiin* liittyviä riippuvuussuhteita.

Esimerkki 9: ROOT/0mark/3advcl/1 ***siirryin***

ja viikko sektiosta **siirryin** jo pikkuhousunsuojiin , **kun** siteet **pysyi** lähes puhtaina .

Esimerkki 10. compound:nn/2nsubj/3ROOT/0 ***tuli***

dpd seurantakoodi tuli melkein seuraavana päivänä , mutta paketti rekisteröityi järjestelmään vasta seuraavana perjantaina .

30 conj

Lähes-luokkaa käsittelevän luvun esimerkeissä 1 ja 2 *conj* esiintyi *advmod*in kanssa samassa biarcissa, ja nämä biarcit olivat yksi lähes-luokan esiin nousevista piirteistä suhteellisen yleisyytensä vuoksi. Kolmessa sen viidestä esiintymästä biarceissa conj siis yhdistyy konjunktiin, mutta muutoin yhdistelmiä on vaikea niputtaa yhteen. *Advmod*-riippuvuussuhdetta ei *melkein*-luokassa esiinny kuin yhdessä biarcissa, eikä se sisällä konjunktia (esimerkki 11.).

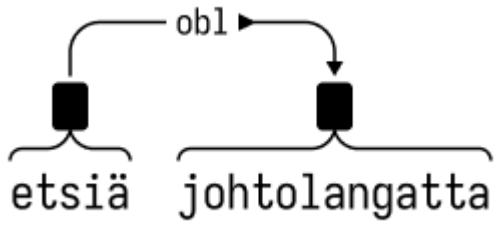
Esimerkki 11: advmod/2amod/3amod/0 ***näköisiä***

tarina etenee hauskasti ; ensin on vain rauhallista ja seesteistä , seuraavaksi **melkein liikkumattoman näköisiä** lehmiä , ja yhtäkkiä kuviin virtaa kissaa ja koiraa " joka nurkasta " ja ne näyttävät todella aktiivisilta .

Esimerkissä 11 huomataan, että tutkimuskohde, sana *melkein,* esiintyy riippvuusssuhteessa *advmod*. Kuten aiemmin mainittua, tämä on sille astemääritteenä luontevaa - adjektiivi, jota se määrittää, luetaan määritesanoihin.

31 Obl

Obl eli obliikvinominaali viittaa sanoihin, jotka eivät ole välttämättömiä tai keskeisiä lauseen ytimen eli vaikkapa juuriverbin kannalta, vaan toimivat sen täydennyksinä. Toiminnallisesti se vastaa verbin, adjektiivin tai toisen adverbin adverbiaalista määritettä. Obliikvinominaali on joko substantiivi, pronomini tai nominaalilauseke, ja suomessa se on usein taivutettuna, tosin se voi esiintyä myös pre- ja postpositioiden kanssa. Obliikvinominaalisuhde voi määrittä esimerkiksi paikkaa, aikaa tai välinettä. (UD) Kuvassa 16 esimerkki (alla).



Kuva 16. Obliikvinominaali kuvattuna lausekkeessa etsiä *johtolangatta*

Obliikvinominaali on erittäin yleinen riippuvuussuhde tarkasteltavissa piirteissämme sekä luokassa lähes että melkein. On kuitenkin huomattava, että luokissa lähes ja melkein obliikvinominaalin kanssa esiintyvät riippuvuussuhteet eivät ole samoja. Kuvan 17 taulukossa näkyy riippuvuussuhteiden jakauma.

lähes	molemmat	melkein
conj	ROOT	obj
nummod		nmod:poss
advmod		nsubj
compound:nn		amod
mark		acl
advcl		
punct		

Kuva 17. Riippuvuussuhteiden jakautuminen yhdistettynä obl-riippuvuuteen.

Esimerkeissä 12 ja 13 oblatiivit täysin erilaisten riippuvuussuhteiden kanssa *lähes*- ja *melkein*-lauseissa.

Esimerkki 12: mark/2advcl/0obl/2 ***otimme***

alkuruokia pääsin tällä kertaa testaamaan lähes koko listan , **koska otimme** vähän kaikkea 9 hlön **ryhmälle** .

Esimerkki 13: amod/2obl/3acl/0 ***olevaa***

olet kestanyt melkein **pahinta** laatuaan **olevaa** kauhukakaraa esimerkillisen hyvin hankalissa **olosuhteissa** .

32 Yhteenveto

Tässä osiossa on tarkasteltu yleisyysperiaatteella valittuja luokkiin lähes tai melkein liittyviä piirteitä (biarceja), joiden kerroin (*coef*) on ei kuin 0 (eli ne määrittävät jompaa kumpaa luokkaa), ja jotka ovat määrittäneen luokkaa sadasta iteraatiosta eli SVM-luokittelijan ajamiskerrasta 90 kertaa tai yli. Kummastakin luokasta analyysin kohteeksi valikoitui 15 piirrettä.

Analyysia varten tarkasteltiin kussakin biarcissa esiintyviä UD-skeeman mukaisia riippuvuussuhdepiirteitä. Luokkaan *lähes* liittyi selvästi *punct*, joka ilmaisee välimerkkejä, ja *conj*, joka ilmaisee rinnakkaisuutta. Lisäksi ainoastaan luokan *lähes* yleisimmissä määrittävissä piirteissä esiintyi riippuvuussuhde *mark*, joka ilmaisee esimerkiksi sivulauseita. Tähän perustuen voisi ajatella, että lauseet, joihin *lähes* liittyy ovat mahdollisesti pidempiä

kuin *melkein*-lauseet, koska rinnasteisuus voi koskettaa listojen lisäksi myös esimerkiksi useampia verbejä kuten rinnasteisissa päälauseissa, välimerkit kuuluvat usein pidempään ilmaisuun ja konjunktiot viittaavat useamman lauseen virkkeisiin. Johtopäätöstä ei voi kuitenkaan tehdä ilman tarkempaa analyysiä ja tilastollisten menetelmien hyödyntämistä.

Luokkaan *melkein* liittyy taas selvästi lauseen subjektia ilmaiseva *nsubj* sekä possessiivista nominaalimääritettä merkitsevä *nmod:poss.* Luokasta nousee esiin myös adjektiivimäärite, eli *amod*, jota ei toisessa luokassa esiinny lainkaan.

Muutamia riippuvuussuhteita vertailtiin luokkien välillä. *ROOT* eli lauseen juuri(verbi) oli yleinen kummassakin luokassa, mutta siihen liittyvät muut suhteet vaihtelivat. Vaikuttaakin siltä, että *melkein*-luokan yleisimmissä piirteissä nousevat esiin lauseiden tyypilliset riippuvuussuhteet, *nsubj* ja *obj*, jotka viittaavat verbin keskeisimpiin täydennöksiin, subjektiin ja objektiin. Luokassa lähes taas oli enemmän vaihtelua, mutta ydinlauseenjäsenet eivät nousseet merkittäviin piirteisiin. Lisäksi tarkasteltiin obliikvinominaalia, eli *obl*-riippuvuussuhdetta. Paitsi juurta *ROOT*, luokat eivät jakaneet yhtään riippuvuussuhdetta suhteessa obliikvinominaaliin.

33 Synonymisuus

Tässä luvussa olen aiemmin esitellyt luokittelijasta saatuja tunnuslukuja. Mallin osuvuus, parhaimmillaan 0,63, on yli kahden luokan luokittelijan perustason 0,5, osoittaa, että tämän aineiston perusteella sanojen *lähes* ja *melkein* välillä on havaittavissa eroja ainakin syntaktisten riippuvuussuhteiden ollessa kyseessä.

Tämä osoittaisi, ettei sanoja lähes ja melkein voi pitää synonyymeinä, kun synonymia määritellään sanojen täydelliseksi korvattavuudeksi toisillaan (esim. Edmonds & Hirst 2002). Sanat eivät kuitenkaan analyysin perusteella eroa toisistaan niin paljon, etteikö niiden välille kannattaisi otaksua selvää samankaltaisuutta. Tällöin melkein-synonymia on hyödyllinen käsite (Edmonds & Hirst 2002).

Tämän tutkimuksen tuottamat tulokset ja analyysi vahvistavat lähtökohtana olevan hypoteesin, että sanojen *lähes* ja *melkein* käytössä on eroja. Ensinnäkin luokittelija pystyy erottamaan luokkia toisistaan. Lisäksi haluaisin nostaa erojen tarkastelusta esiin ne riippuvuussuhteet, jossa luokkiin liittyvissä biarceissa on yhteinen, melko yleinen tekijä (*ROOT* tai *obl*), mutta joiden esiintymisyhteydet toisten riippuvuussuhteiden kanssa ovat varsin erilaiset.

Koska sanojen riippuvuussuhdeprofiilit ovat vertailun perustana, voi myös tulkita, että sanojen käyttö määrittyy ainakin joiltakin osin nimenomaan lauserakenteen pohjalta. Toisaalta syntaktiset riippuvuussuhteetkin kantavat mukanaan informaatiota sanastosta, joten mahdollisesti riippuvuussuhdeprofiilienkin kautta voidaan tehdä päätelmiä myös astemääritteiden sanastollisista konteksteista.

Kenties vahvempia tuloksia voisi olla saatavilla, kun tukivektorikoneen hyperparametrejä lähdettäisiin tutkimaan tarkemmin. Lisäksi piirteiden analysoimiseen olisi hyödyllistä saada systemaattinen työkalu. Voidaan kuitenkin todeta vahvistuneen olettaman, että melkein ja lähes ovat melkein-synonyymeja.

34 Johtopäätökset

Tämä tutkimus pyrki vahvistamaan teoreettisen, kielitajuun perustuvan olettaman, jonka mukaan sanat *melkein* ja *lähes* eivät ole synonyymeja, vaan melkein-synonyymeja. Tulosten pohjalta voidaan todeta, että sanojen välille on löydettävissä eroja, jotka vaikuttavat niiden käyttöön. Kognitiivisen kielioppiteorian mukaan sanojen merkitys liittyy niiden käyttöyhteyksiin, joten mikäli käyttöyhteyksissä on eroja, on sanojenkin välille syytä olettaa merkitysero. On intuitiivisesti ja tämän tutkimuksen tuloksiin perustuen selvää, että erot ilmaisujen välillä ovat hyvin hienovaraisia.

Tässä tutkimuksessa näitä hienovaraisia eroja on tarkasteltu sanojen riippuvuussuhdeprofiileihin pohjautuen. Riippuvuussuhdeprofiilit muodostuvat sanojen käyttöyhteyksistä, ja ne voidaan luoda Universal Dependencies -skeeman mukaan annotoidulla datalla, josta sanaluokkia koskeva informaatio on poistettu. Tämä tarkastelukulma sopii myös kognitiiviseen käsitykseen kielestä; riippuvuussuhdeprofiilit ovat puhtaasti syntaktisia ja ilmaisevat semanttisen kontekstin sijaan syntaktisia yhteyksiä. Lisäksi UD-skeema on hyödyllinen jäsennystapa, kun halutaan laajentaa riippuvuuussuhdeprofiilien käyttöä myös tämän tutkimuksen kohdekielen ulkopuolelle.

Jotta riippuvuussuhteita voitaisiin tarkastella monipuolisemmin, olisi hyödyllistä kirjoittaa ohjelma, joka yhdistäisi analyysistä saadut piirrebiarcit suoraan lauseessa tarkoitettuihin yhteyksiin, eli nostaisi leksikaalisesta datasta sen ilmaisun, jota biarc edustaa. Tämä palauttaisi riippuvuussuhteet myös semanttiselle tasolle, joka voisi osoittaa esimerkiksi tietynlaista sanastollista jakoa sanojen välille.

Tämä tutkimus on ollut vain pintaraapaisu melkein-synonymian korpuslähtöiseen tutkimukseen. On helppo kuvitella, mitä vastaavilla tutkimusmetodeilla voitaisiin tarkastella. Metodiikkaa ja työkalua, luokittelijaa ja riippuvuussuhdeprofiilia, voitaisiin soveltaa melko suoraan muiden melkein-synonyymien tutkimiseen. Aineistona voitaisiin käyttää monenlaista UD-skeemasta saatavaa informaatiota tai lauseita sellaisenaan. Myös erilaisten korpusten käyttäminen voi joissakin tutkimuksissa olla mielekästä. Finnish Internet Parsebank tarjoaa merkittävän määrän spontaanisti suomeksi tuotettua kieltä oikeilta kielenkäyttäjiltä, vaikka se ei esimerkiksi erottele äidinkielisen suomentuottajan, suomea toisena kielenä puhuvan ja konekäännöksen tuottamaa kieltä.

Kun erilaisia käyttöyhteyksiä pystytään tarkastelemaan jopa pienten erojen ollessa kyseessä, tai kun sanojen merkityksiä pystytään mallintamaan niiden ympäristön laskennalliseen dataan

perustuen, voiko sillä olla myös vaikutus kognitiivisen kielioppiteorian kehitykseen? Loppujen lopuksi, mikäli kieltä ei ole syytä irrottaa sen käyttöyhteydestä, onko kielioppiteoriaa syytä irrottaa sen pohjana olevasta ilmiöstä, kielenkäytöstä ja kielellisistä riippuvuussuhteista? Voiko olla löydettävissä sellainen spontaanisti tuotetun kieliaineksen mallintamisen keino, joka toisi meidän kielenymmärrykseemme jotakin sellaista, jota emme ole osanneet edes kuvitella?

Lähteet

Arppe, Antti. *Univariate, Bivariate and Multivariate Methods in Corpus-Based Lexicography : A Study of Synonymy.* Helsinki: University of Helsinki, 2008. Print.

Bolinger, Dwight. Entailment and meaning of structures. *Glossa* 2 (1968): 119-127.

Cruse, D. A. *Lexical Semantics.* Cambridge: Cambridge University Press, 1986. Print.

Divjak, Dagmar, and Stefan Th Gries. Clusters in the Mind?: Converging Evidence from near Synonymy in Russian. *The mental lexicon* 3.2 (2008): 188–213. Web.

Edmonds, Philip, and Graeme Hirst. Near-Synonymy and Lexical Choice. *Computational linguistics - Association for Computational Linguistics* 28.2 (2002): 105–144.

FIP = Finnish Internet Parsebank https://turkunlp.org/finnish_nlp.html viitattu 30.12.2023

Hays, David. Dependency Theory: A Formalism and Some Observations. *Language* 40.4.(1964): 511–525.

Huumo, Tuomas. Skalaarisuuden kielioppia. Kvanttoreiden ja astemääritteiden keskinäiset määrityssuhteet. *Sananjalka* 64 (2022), 86-107.

On the Gradable Nature of the Search Domain: A Study of Degree Modifiers and the Scalar Semantics of Finnish Spatial Grams. *Cognitive Semantics (online)* 2021.1 (2021): 85–113. Web.

Jantunen, Jarmo Harri. *Synonymia ja käännössuomi : korpusnäkökulma samamerkityksisyyden kontekstuaalisuuteen ja käännöskielen leksikaalisiin erityispiirteisiin.* Joensuu: Joensuun yliopisto, 2004. Print.

KTS = Kielitoimiston sanakirja. 2021. Helsinki: *Kotimaisten kielten keskus* Verkkojulkaisu HTML. Päivitettävä julkaisu. Päivitetty 11.11.2021. Viitattu 1.12.2023

Laippala, Veronika et al. "Dependency Profiles in the Large-Scale Analysis of Discourse Connectives." *Corpus linguistics and linguistic theory* 17.1 (2021): 143–175. Web.

Langacker, Ronald W. *Cognitive Grammar : A Basic Introduction.* Oxford: Oxford University Press, 2008. Print.

Luotolahti, Juhani et al. Towards Universal Web Parsebanks. *Proceedings of the International Conference on Dependency Linguistics (Depling 2015) (*2015): 211-220. Web.

de Marneffe, Marie-Catherine et al. 2014. Universal Stanford dependencies: A cross-linguistic typology. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (2014), pages 4585–4592.

Paquot, Magali, and Stefan Th Gries. A Practical Handbook of Corpus Linguistics. Cham: Springer International Publishing AG, 2021. Web.

Paradis, Carita. Adjectives and boundedness. *Cognitive Linguistics* 12 (2001), 47–64.

Pedregosa et al. Scikit-learn: Machine Learning in Python. *JMLR* 12 (2011): 2825-2830.

SES = Suomen etymologinen sanakirja. 2022. Helsinki: Kotimaisten kielten keskus. Verkkojulkaisu HTML. Päivitettävä julkaisu. Päivitetty 5.12.2023. https://kaino.kotus.fi/ses/?p=article&etym_id=ETYM_d5a16ac965a24eef5f0e7fcc1f527a62&word=melkein Viitattu 28.12.2023

https://kaino.kotus.fi/ses/?p=article&etym_id=ETYM_4e4bedbe1b139d62d6ca398af9865674&word=l%C3%A4hes Viitattu 28.12.2023

TNPP = Turku Neural Parser Pipeline, verkkosivu https://turkunlp.org/Turku-neural-parser-pipeline/ viitattu 30.12.2023

UD = Universal Dependencies, verkkosivu https://universaldependencies.org viitattu 20.12.2023; ks. myös de Marneffe.

Vanhatalo, Ulla. “Kyselytekstit vs. korpuslingvistiikka lähisynonyymien semanttisten sisältöjen arvioinnissa-mitä vielä keskeisestä ja tärkeästä?” Virittäjä 107.3 (2003): 351–351. Print.

VISK = Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen ja Irja Alho 2004: Iso suomen kielioppi. Helsinki: Suomalaisen Kirjallisuuden Seura.

Liitteet

Liite 1. Koodi

Tukivektorikoneen koodi osoitteessa https://github.com/petraimil/progradu_svm

Liite 2. Aineisto

Aineisto nähtävillä osoitteessa http://dl.turkunlp.org/biarc_data/datafiles/