hyväksymispäivä arvosana

arvostelija

Ironisten ilmaisujen automaattinen tuottaminen suomeksi

Juhani Bergström

Helsinki 21.5.2015

Pro gradu -tutkielma

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen tiedekunta	Tietojenkäsittelytieteen laitos

Tekijä — Författare — Author
Juhani Bergström

Työn nimi — Arbetets titel — Title

Ironisten ilmaisujen automaattinen tuottaminen suomeksi

Oppiaine — Läroämne — Subject
Tietojenkäsittelytiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	21.5.2015	46 sivua + 0 liitesivua

Tiivistelmä — Referat — Abstract

Tässä työssä sovelletaan Tony Vealen aiempaa englanninkielistä ironisten vertausten tuottamista käsittelevää tutkimusta ja sen menetelmiä suomen kieleen toteuttamalla prototyyppijärjestelmä. Järjestelmällä on tarkoitus tuottaa ironisia ilmauksia jonkin halutun sanan pohjalta, käyttäen annettuja strategioita eli sääntöjä. Toteutuksessa yritetään ensin tunnistaa aineistona olevasta tekstistä sanojen perusmuodot ja niiden tyypit. Tämän jälkeen tekstistä yritetään löytää erilaisilla malleilla sanojen välisiä sanayhteyksiä eli sanoihin liittyviä stereotyyppisiä ominaisuuksia sekä sanoihin läheisesti liittyviä naapurisanoja. Kun sanayhteydet on löydetty voidaan annetusta sanasta johtaa strategioiden avulla uusia, toivottavasti ironisia, ilmauksia. Strategioilla uusi ilmaus johdetaan annetusta sanasta käyttämällä sanayhteysoperaattoreita. Operaattorilla voidaan tuottaa esimerkiksi sanaan läheisesti liittyvä sana, sanan vastakohta tai sanaa ominaisesti kuvaava sana. Operaattoreita voidaan myös yhdistää, jolloin voidaan tuottaa esimerkiksi sanaan läheisesti liittyvän sanan vastakohta. Menetelmä vaikuttaa lupaavalta uusien ilmausten muodostamiseen, mutta ironian tuottaminen osoittautui hyvin hankalaksi.

ACM Computing Classification System (CCS):
Computing methodologies -> Natural language generation

Avainsanat — Nyckelord — Keywords
Luonnollisen kielen käsittely

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Sisältö

1 Johdanto

Tässä tutkielmassa on tarkoitus soveltaa Tony Vealen tutkimuksessaan ”Detecting and Generating Ironic Comparisons: An Application of Creative Information Retrieval” [Vea12] käyttämiä menetelmiä suomen kieleen ja samalla selvittää menetelmien toimivuus suomen kielelle, sekä tutkia miten englannin ja suomen kielen rakenteelliset erot vaikuttavat menetelmien toimivuuteen. Soveltuvuus selvitetään toteuttamalla prototyyppijärjestelmä Vealen käyttämillä menetelmillä.

Tutkimuksessaan Veale on yrittänyt luoda luovia ja ironisia vertauksia (engl. simile), jotka ovat pohjimmiltaan muotoa ”nopea kuin gepardi”. Tämä on tehty luomalla ensin tietokanta tunnetuista vertauksista ja sen jälkeen on erilaisia sääntöjä ja sanayhteyksiä käyttämällä luotu näistä uusia vertauksia. Tekniikalla on onnistuttu luomaan monia lupaavia vertauksia, mutta niiden ironisuus on kyseenalaista ja ironisten vertausten tuottaminen vaatii lisätyötä.

Keskeisinä osina toteutuksessa ovat sanojen välisten yhteyksien tunnistaminen aineistosta ja uusien vertausten muodostaminen näiden yhteyksien pohjalta. Aineistona, josta sanayhteydet kerätään, käytetään suomenkielistä Wikipediaa. Sanojen välisiä yhteyksiä muodostetaan tutkimalla tekstiaineistosta muun muassa sitä, mitkä adjektiivit esiintyvät usein tietyn substantiivin yhteydessä ja toisinpäin, sekä mitkä adjektiivit/substantiivit esiintyvät usein keskenään lähekkäin. Näin saadaan aikaan yhteyksiä adjektiivien ja substantiivien välille sekä keskenään samantyyppisten adjektiivien/substantiivien välille. Näiden lisäksi käytetään aineistoa sanojen vastakohdista.

Kerättyjä sanojen välisiä yhteyksiä käytetään muodostamaan uusia vertauksia tietyn ominaisuuden/sanan pohjalta erilaisten strategioiden avulla. Esimerkiksi ottamalla ominaisuuden, tässä tapauksessa sana kylmä, läheisyydessä usein esiintyvä sana ja ominaisuuden kanssa usein esiintyvä sana, saadaan johdettua vertaus kylmä kuin märkä kala. Edellisellä strategialla saadaan myös useita muita vertauksia kuten kylmä kuin sydämetön robotti.

Suomen kielellä tarjolla olevien valmiiden aineistojen saatavuus ja rajallisuus verrattuna englanninkielisiin aineistoihin saattaa aiheuttaa rajoitteita joidenkin osien toteutuksessa. Myös sanojen taivutusmuodot, kaksiosaiset sanat ja yhdyssanat ja niiden tulkitseminen voivat olla ongelmallisia, esim. muoviveitsi vai muovinen veitsi.

Luvussa 2 esitellään aiempaa tutkimusta ja Vealen järjestelmä, jota tässä tutkimuk-

sessa sovelletaan. Luvussa 3 esitellään itse toteutettu järjestelmä ja tarkastellaan sen toimintaa. Luvussa 4 tarkastellaan toteutuksella saatuja tuloksia.

2 Aiempi tutkimus

Tässä luvussa kuvataan ensin tämän tutkielman aiheeseen liittyvää aiempaa tutkimusta yleisemmin ja sitten Vealen käyttämä ja tässä tutkielmassa sovellettava menetelmä yksityiskohtaisemmin.

2.1 Yleistä

Aiempaa tutkimusta tästä aiheesta on melko vähän ja sitä on myös hankala löytää koska ei ole mitään tiettyä yleistä vakiintunutta nimeä kuvaamaan aihetta ja tätä kyseistä menetelmää. Veale itse kutsuu menetelmäänsä nimellä Creative Information Retrieval (CIR).

Vealen mukaan [Vea11] merkittävin ero tällä CIR-menetelmällä aiempiin tutkimuksiin on se, että tässä yhdistetään kuvaannollisen kielen käsittelyn (figurative language processing, FLP) ja tiedonhaun (information retrieval, IR) menetelmiä. IR ja FLP käsittelevät kieltä ja merkitystä hyvin eri lailla. IR käsittelee kieltä avoimena joukkona merkkejä, joiden mukaan tekstiä voidaan indeksoida ja hakea. Se keskittyy tekstin merkityksellisyyteen eikä niinkään sen merkitykseen/tarkoitukseen. FLP sen sijaan käsittelee kieltä epävakaiden merkkien järjestelmänä, joita voidaan käyttää uusilla luovilla tavoilla. Erona on myös se, että IR on käytännöllinen, skaalautuva ja jokapäiväisessä käytössä. FLP ei sen sijaan ole skaalautuva eikä tarpeeksi käytännöllinen yleiseen käyttöön. Mutta Vealen esittämä CIR, joka on IR:n ja FLP:n sekoitus, tuo tiedonhaun menetelmien käyttöön uusia operaattoreita, jotka mahdollistavat ei-kirjaimellisen ilmausten haun ja saavat kuvaannollisen kielen käsittelyn skaalautuvammaksi. Tämän hybridin rungon päälle voidaan rakentaa käytännöllisiä kielellisesti luovia sovelluksia.

IR on tehokas mekanismi tekstin hallintaan ja se on laajasti käytössä. Mutta se ei ole erityisen luova mekanismi uusien käsiterakenteiden muodostamisessa ja vanhojen uudelleenorganisoimisessa uusin tavoin ja kekseliäästi käyttäjän hakemaa tietoa kuvaavien dokumenttien havaitsemiseksi [Vea04]. Koska kieli on dynaaminen ja hyvin luova ilmaisukeino, haettavat käsitteet ovat liikkuva kohde IR-järjestelmälle. Veale esittää että vain ajattelemalla luovasti voi IR-järjestelmä noutaa tehokkaasti

dokumentteja, jotka sisältävät luovaa kieltä.

Tiedonhaussa käyttäjän valitsemat hakusanat kuvaavat hänen informaation tarvettaan, mutteivät välttämättä kuvaa parasta sanajoukkoa, jolla halutun informaation saisi haettua aineistosta. Sen sijaan että hakusanoja käytettäisiin suorana hakuna, älykkäät hakukoneet käyttävät niitä itse haun muodostamiseen. Tämä haun laajennus (query expansion) yrittää muodostaa monipuolisen haun käyttäjän antamista hakusanoista ja näin löytää useamman varteenotettavan dokumentin. Haun laajentaminen voidaan toteuttaa useilla tavoilla, joista jotkut ovat hyvin yksinkertaisia, mutta vain osaa voidaan pitää luovina [Vea04]. Tilastolliset tekniikat voivat tunnistaa aiheeseen liittyviä yhteyksiä ilmauksien välillä ja näin hakua voidaan laajentaa hakusanoilla, jotka liittyvät samaan aiheeseen. Esimerkiksi tekstiaineiston analyysi (corpus analysis) paljastaa vahvan esiintymisyhteyden sanapareille lääkäri ja hoitaja, sairaala ja terveydenhuolto, Jaguar ja urheiluauto. Näitä tietoja hyödyntämällä voidaan muodostaa aihetta paremmin kuvaava laajennettu haku. Käyttäjän palautetta hyödyntävät tekniikat (relevance feedback) puolestaan laajentavat hakua käyttämällä ilmauksia, joita löytyy dokumenteista, jotka käyttäjä on merkinnyt liittyvän aiheeseen. Tarkoitus on löytää enemmän dokumentteja, jotka ovat samanlaisia kuin käyttäjän merkitsemät ja eroavat niistä, joita käyttäjä ei valinnut. Samoin toimivat ilman käyttäjän palautetta LCA-menetelmät (Local context analysis), jotka käyttävät käyttäjän hakusanoja ensimmäisenä hakuna ja sitten etsivät parhaista tuloksista tilastollisella analyysillä lisää sanoja ja ilmauksia uuteen hakuun, jolla haetaan lisää dokumentteja. Tilastolliset tekniikat eivät tyypillisesti tarvitse tietämystä aihealueesta ja useat pystyvät sopeutumaan erilaiseen sanojenkäyttöön automaattisesti, jos niitä harjoitetaan, kun uusia dokumentteja ilmaantuu aineistoon. Asiantuntijajärjestelmät (knowledge-based techniques) käyttävät mallia aihealueesta tunnistaakseen haetun asian käyttäjän hakusanoista ja näin niihin liittyviä määritelmiä voidaan hyödyntää. Malli voidaan saada yleiskäyttöisestä leksikaalisesta ontologiasta, kuten WordNetistä, joka on laaja-alainen jäsennelty sanasto. Haun laajennusprosessi hyödyntää tietoa synonyymeistä, ylä- ja alakäsitteistä, sekä siitä onko käsite toisen osa (esim. sormi-käsi). Nämä tekniikat kuitenkin tulkitsevat kyselyn ja sen käsitteellisen sisällön hyvin kirjaimellisesti. Jotkin laajat ontologiat sisältävät tietoa metaforista ja käsitteellisistä rakenteista, jolloin ne pystyvät ymmärtämään näiden merkityksiä. Koska tuoreus on tavoiteltavaa luovassa ajattelussa, tilastolliset tekniikat eivät löydä riittävästi tietoa, josta saisi johdettua yhteydet luovien metaforien ymmärtämiseksi. Vaikkakin joidenkin vakiintuneiden tavanomaisten metaforien tilastollisessa ymmärtämisessä on ollut onnistumisia.

Tiedonhaussa (IR) perusolettamuksena on että käyttäjä pystyy muuttamaan informaation tarpeensa tehokkaaksi hauksi käyttämällä samanlaista kieltä, jota aiheesta käytetään haun kohdeaineistossa. Jos kokoelma sisältää luovempia tai mielikuvituksellisempia ilmauksia aiheesta, niin myös haun pitää olla muotoiltu samoin. Veale esitteli idean luovasta informaation hausta (creative information retrieval) tutkiessaan kuinka IR-järjestelmä voi itse tuottaa osan luovasta ennakkoaavistuksesta toimien välittäjänä kirjaimellisen merkityksen ja luovien merkitysten välillä [Vea04]. Ennakkoaavistuksella tarkoitetaan tässä sitä, että järjestelmän pitää osata aavistaa millä sanoilla mistäkin asiasta puhutaan, esimerkiksi käytetäänkö sarjakuvakirjasta muotoa ”comic book” vaiko muotoa ”graphic novel”. Näin käyttäjän ei tarvitse osata käyttää muodikkaita hakusanoja hauissa. Tällainen ennakkoaavistus voi vaihdella yksinkertaisesta uudelleen muodostamisesta leikkisiin vihjauksiin ja uudelleen määrityksiin. Teksti voi esimerkiksi käsitellä Koraania, vaikka se sisältäisi vain ilmauksen ”muslimi raamattu” tai kumiyhtiön johtajaa voidaan kuvata leikillisesti ”kumiparoniksi”. Luova IR-järjestelmä voi jopa osata muodostaa sanoja sanakirjan ulkopuolelta, kuten esimerkiksi ”chocoholic” ja ”sexoholic”. Tavanomaiset IR järjestelmät käyttävät erilaisia haun laajennustekniikoita, jotka automaattisesti vahvistavat käyttäjän hakua lisäämällä avainsanoja tai painotuksia, jotta olisi mahdollista osua asiaankuuluviin teksteihin, joita ei muuten löytyisi. Tekniikat vaihtelevat stemmereiden ja morfologisten analyysien käytöstä sanastojen (kuten WordNet) hyödyntämiseen synonyymien lisäämiseksi, tilastollisten analyysien käyttöä tapauskohtaisten sanayhteyksien ja miltei synonyymien löytämiseksi. Jotkin tekniikat voivat ehdottaa tavanomaisia metaforia, jotka ovat päätyneet sanakirjaan, ne tuskin kuitenkaan pystyvät tunnistamaan suhteellisen uusia ilmaisuja.

Järjestelmän luovuttaa voidaan testata sen kyvyllä tehdä olettamuksia ja luoda uusia käsitteitä tilanteen mukaisesti [Vea04]. Tällainen testi erottelee ne järjestelmät, jotka jäljittelevät luovuutta ennalta määriteltyjen sääntöjen mukaan ja hakutaulujen (look-up table) avulla, niistä jotka osoittavat aitoa kekseliäisyyttä. Esimerkiksi järjestelmä voi näyttää ymmärtävän metaforia, jos sillä on riittävän suuri sanakirja ja sopivia poikkialaisia assosiaatioita, mutta sellaista järjestelmää ei voi sanoa luovaksi, sillä se rajoittuu vain käsitteisiin, jotka ovat entuudestaan järjestelmän tietokannassa. Sen sijaan järjestelmä, joka tuottaa yhteydet itse lennosta, ja joka ei siksi tukeudu vain ennalta määriteltyyn tietoon, on luova kun tarkoitetaan uuden tietämyksen luomista.

Aiemmin Veale on käyttänyt tämän CIR-menetelmän periaatteita pohjana juurikin työssään hakujen laajentamiseksi [Vea04]. Keskeisenä tavoitteena oli muodostaa

luovia hakuja, joiden tavoite on ymmärtää merkitys pelkkien sanojen ja synonyymien sijaan ja näin hauilla onnistutaan löytämään myös eri sanoilla esitetyn saman asian. Esitelty menetelmä keskittyy käyttämään WordNettiä ja laajentamaan hakusanoja sen avulla. Menetelmä hyödyntää WordNetin sanojen yhteyksien tietoja (synonyymit, eri merkitykset, sanan ylä- ja alakäsitteet), metaforia ja sanojen määrityksiä/selityksiä, mutta laajentaen niitäkin lisää sanojen yhteyksien avulla. Tämä tehdään kuitenkin niin, että haku pysyy aiheessa, eikä hakualue leviä kovin laajaksi. Esimerkiksi käyttäjän tekemä haku ”Italian recipes”, muuttuu päättelyketjun läpikäymisellä hauksi: ”((Italian and (cookery or food)) or pizza or antipasto or Mozzarella or...) near (book or cookbook or recipes or bible))” Tämä laajennettu haku mahdollistaa osumisen luovasti vihjaileviin teksteihin kuten ”The Antipasto Bible”, jotka sisältävät vahvan yhteyden hakuun ”Italian recipes”.

2.2 Aineiston/sanayhteyksien keruu

Tässä tutkielmassa käsiteltävän työn aiheeseen vahvasti liittyen Veale on tutkinut myös ironisten vertausten tunnistamista automaattisesti [VeH10]. Tutkimus tuotti sääntöpohjaisen järjestelmän ja laajan aineiston ironisiksi ja ei ironisiksi luokiteltuja vertauksia. Järjestelmä käytti sekä normaaleja että ironisia vertauksia ja yritti selvittää voiko vertauksen ”as X as Y” esittää hieman muunnellussa muodossa, joka on hyvin todennäköisesti ironiaton vai onko se todennäköisemmin esitettävissä muodossa, joka on hyvin todennäköisesti ironinen. Todennäköisyydet pääteltiin pääosin Googlen avulla selvitetyn web-esiintymistiheyden perusteella, mutta myös muita päättelysääntöjä käytettiin, säännöt pohjautuivat englanninkielisissä ironisissa vertauksissa havaittuihin ominaispiirteisiin. Esimerkiksi sääntöinä käytettiin sitä esiintyykö vertaus useammin ”about” sanan kanssa vai ilman internetissä ja ovatko vertauksen sanat morfologisesti samanlaisia keskenään kuten ”as masculine/manly as a man”. Tulokset olivat lupaavia ja järjestelmä onnistui tunnistamaan 87% ironisista vertauksista ja 89% normaaleista vertauksista.

Vertauksia ironian tunnistamisen testijoukkoon kerättiin käyttämällä Google-hakuja muotoa ”about as * as *”. Vertaukseen on otettu WordNetistä adjektiivipareja jolloin haut ovat olleet esimerkiksi ”about as strong as *” ja ”about as weak as *”. Tuloksena saatiin noin 45 tuhatta vertausta, joissa kuitenkin on usein pronomini viittaamassa ympäröivään tekstiin. Siksi aineisto käytiin manuaalisesti läpi ja saatiin noin 20 tuhatta erilaista vertausta, joista noin 15 tuhatta oli ironisia. Muotoa ”about” käytettiin, koska se löytää hyvin ironisia vertauksia. Veale on aiemmin käyt-

tänyt vertausten muotoa ”as X as Y” (esimerkiksi ”as hot as an oven”) yleisten vertausten ja niiden sisältämien stereotyyppien keräämisessä internetistä. Mutta tässä Veale ja Hao näyttävät että malli ”about as X as Y” löytää yhtä suuren joukon luovia, enimmäkseen ironisia, vertauksia. Samalla he myös osoittavat että suuri sanasto stereotyyppisiä käsitteitä (yli 4000 substantiivia) ja niiden keskeisiä ominaisuuksia (yli 2000 adjektiivia) voidaan kerätä internetistä.

Veale kertoo, että aiemmista tutkimuksista selviää, että muotoa ”Xs and other Ys” olevaa mallia voidaan käyttää luovempien tapauskohtaisten luokittelujen muodostamiseen kuin suoraan WordNetistä poimittaessa [Vea11]. Esimerkiksi ilmaus ”athletes and other celebrities” vihjaa yhteydestä, jossa urheilijat nähdään tähtinä. On myös näytetty kuinka ilmauksista muotoa ”Xs like A, B and C” voidaan johtaa joustavia luokitteluja. Lisäksi on näytetty kuinka tilastollisella analyysillä voidaan kokoelmasta tunnistaa ja kerätä automaattisesti tavanomaisia metaforia. Luovemmat metaforat sen sijaan ovat haasteellisempia [Vea11].

Veale ja Hao tutkivat aiemmin vertausten käyttöä stereotyyppisten normien lähteenä ja keräsivät internetistä kymmeniä tuhansia vertauksia, jotka ovat muotoa ”as X as a Y”. Koska vertausten ironisuuden tunnistaminen koneellisesti luotettavasti on mahdotonta, he lajittelivat manuaalisesti vertaukset normaaleihin ja ironisiin. Lopputuloksena oli yli 12000 normaalia vertausta, kuten ”as hot as an oven”, ja lähes 3000 ironista vertausta, kuten ”as subtle as a sledgehammer”. Tämän jälkeen he käyttivät kerättyjä normaaleja vertauksia stereotyyppisten normien (esimerkiksi; uunit ovat kuumia, viidakot kosteita ja lumi pehmeää) lähteenä systeemissään metaforien ymmärtämiseksi ja tuottamiseksi [Vea12] .

2.3 CIR-menetelmä

Tässä luvussa esitellään CIR-menetelmä, jonka Veale on esitellyt tutkimuksessaan Detecting and Generating Ironic Comparisons: An Application of Creative Information Retrieval [Vea12]. Ensin etsitään sanayhteyksiä ja sitten näiden yhteyksien pohjalta muodostetaan uusia vertauksia erilaisilla strategioilla.

2.3.1 Sanayhteyksien kerääminen

Googlen n-grammeista löydettyjä sanojen välisiä yhteyksiä käytetään hyväksi muodostettaessa uusia vertauksia. Nämä n-grammit ovat internetin teksteistä poimittuja n sanan mittaisia tekstinosia. Kuhunkin tekstinosaan liittyy myös tieto sen koko-

naisesiintymismäärästä. Aiemmassa tutkimuksessa stereotyyppisten ominaisuuksien keräämiseen käytetty malli ”as X as Y” oli kovin rajoittunut, eikä sillä saatu kaikkia adjektiivien ominaisuuksia kerättyä. Se ei tunnistanut esimerkiksi, että vauvat kuolaavat, poliitikot valehtelevat ja koirat haukkuvat. Kerättyjen ominaisuuksien määrän kasvattamiseksi ja stereotyyppisten ominaisuuksien tunnistamiseksi paremmin käytetään seuraavanlaista mallia. Aluksi on kerätty Googlen 3-grammeista kaikki muotoa <DET PROPERTY NOUN> olevat sanonnat. Tässä DET on määrittäjä (esimerkiksi the, a[n], this, that, my, their, many, few ja several). PROPERTY voi olla joko adjektiivi WordNet:istä tai taivutettu verbi, joka osoittaa käyttäytymistä, kuten ”swaggering” tai ”armored”. Tapauksista, joissa PROPERTY-sana on adjektiivi, muodostetaan ”as”-vertaus ”as ADJ as a NOUN”, kun taas toiminnallisille sanoille käytetään ”like”-vertausta muodossa ”BEHAVIOR like a NOUN”. Näin Googlen 3-gramista ”a reckless cowboy” tuotetaan vertaus ”as reckless as a cowboy” ja 3-gramista ”a swaggering cowboy” muodostetaan vertaus ”swaggering like a cowboy”.

Tämän jälkeen muodostettuja vertauksia käytetään web-hakuina ja näin saadaan selville vertauksen yleisyys verkossa, mistä päätellään se, onko vertauksen sanoilla keskinäinen yhteys. Vain vertaukset, jotka saavat osumia, hyväksytään yhteydeksi sanojen välille. Näin muodostettuun joukkoon kuuluu kuitenkin monia epämääräisiä vähän kuvaavia ilmauksia kuten ”walking like a drunk” tai ”talking like a baby”. Näistä voisi erotella tilastollisilla menetelmillä edellisten kaltaiset vähiten informatiiviset kuvaavimmista kuten ”staggering like a drunk” ja ”babbling like a baby”. Koska työ pitää kuitenkin tehdä vain kerran ja tuloksena saadaan tarkka käsite aineisto Veale ja Hao suorittivat työn manuaalisesti muutamassa viikossa. Saatu stereotyyppimalli on huomattavasti laajempi kuin aiemmin tuotettu. Se pitää sisällään 9479 eri stereotyyppiä ja jokaiseen liittyy valikoima 7898:sta eri ominaisuudesta. Kaiken kaikkiaan malli käsittää yli 75000 yksilöllistä substantiivi-ominaisuus yhteyttä. Tämä on huomattavasti enemmän kuin saadut reilut 12000 yhteyttä aiemmasta tutkimuksesta. Esimerkiksi sanaan ”baby” liittyy 163 stereotyyppistä ominaisuutta.

2.3.2 Uusien vertausten muodostus

Vertauksia muodostetaan erilaisten strategioiden/sääntöjen avulla, joissa sanojen välisiä yhteyksiä kuvaavat operaattorit on yhdistelty muodostamaan uusi vertaus. Näitä sanojen välisiä yhteyksiä kuvaavat operaattorit esitetään merkeillä @, ? ja ˆ ja ne toimivat seuraavanlaisesti:

Operaattori **@** on stereotyyppioperaattori, joka liittää substantiivin ja sitä kuvaavan adjektiivin toisiinsa. Esimerkiksi @partaveitsi voisi liittyä sanoihin terävä ja sileä kun taas @terävä voisi liittyä sanoihin partaveitsi, puukko ja miekka.

Operaattori **?** on naapurusto-operaattori, joka liittää substantiivin toiseen substantiiviin, jotka esiintyvät usein tekstissä lähekkäin. Samoin liitetään lähekkäin esiintyvät adjektiivit toisiinsa. Substantiivien välinen yhteys havaitaan ”ja”-sanan avulla esimerkiksi ”enkelit ja demonit” tai ”lääkärit ja hoitajat” tapaisista tekstin osista, joissa molemmat sanat ovat monikossa. Esimerkiksi ?disaster liittyisi sanoihin ”tragedy”, ”catastrophe”, ”calamity”, ”misfortune”, ”hardship”, ”plague”, ”famine” ja niin edelleen. Adjektiivien välinen yhteys sen sijaan havaitaan muotoa ”yhtä X ja Y kuin” (engl. ”as X and Y as”) olevista tekstin osista. Esimerkiksi ?tragic liittyisi sanoihin ”sad”, ”shocking”, ”terrible”, ”unfortunate”, ”ridiculous” ja niin edelleen. Substantiivin väliset yhteydet naapureihinsa on järjestetty niiden WordNet-yhtäläisyyden mukaan. Adjektiivin yhteydet naapureihinsa on järjestetty niiden esiintymistiheyden mukaan.

Operaattori ^ on kategoriaoperaattori, jolla ^luokka liittyy kaikkiin ennalta määritetyn luokan jäseniin. Luokkia voi luoda esimerkiksi tekemällä haun @terävä & ^työkalu ja näin muodostettaisiin luokka ”terävätyökalu”. Tässä esimerkissä siis olisi jo olemassa luokka työkalu. Operaattoria voidaan käyttää myös viittaamaan WordNet luokkiin. Esimerkiksi ^henkilö viittaisi kaikkiin henkilöä ilmaiseviin substantiiveihin WordNetissä.

Näiden lisäksi on vielä yksi operaattori.

Operaattori **-** on vastakohtaoperaattori, joka liittää adjektiivin vastakohtaansa. Vastakohdat on saatu selville WordNetistä. Näin esimerkiksi sana pehmeä liittyy sanaan kova ja vahva heikkoon.

Näitä edellä mainittuja operaattoreita voidaan yhdistää. Esimerkiksi ?@adj liittyy mihin tahansa substantiiviin, joka on naapurustossa sellaiselle sanalle, joka kuvaa annettua adjektiivia. Sen sijaan @-adj tuottaa minkä tahansa annetun adjektiivin vastakohdalle kuvaavan substantiivin. Vealen mukaan ilmeisimmät ja käytännöllisimmät operaattoreiden yhdistelmät ovat seuraavat:[Vea11]

Naapureiden naapurit ?? kun ?X löytää kaikki X:n naapurit, niin ??X tuottaa kaikki näiden naapureiden naapurit. Naapurin naapuri liittyy kaikkiin sanan naapureiden naapureihin, eikä vain yhden naapurin kaikkiin naapureihin. Tällöin esimerkiksi ??artist liittyy useampiin sanoihin kuin ?artist tuottaen enemmän monimuotoisuutta, enemmän kohinaa (noise) ja enemmän luovia

mahdollisuuksia.

Stereotyyppien stereotyypit @@ löytää kaikkien stereotyyppien stereotyypit. Esimerkiksi @@diamond liittyy kaikkiin substantiiveihin, jotka jakavat minkä tahansa stereotyyppisen adjektiivin sanan ”diamond” kanssa. Samoin @@sharp liittyy kaikkiin kaikkien niiden substantiivien stereotyyppisiin ominaisuuksiin, joille ”sharp” on ominainen stereotyyppi.

Stereotyyppien naapurit ?@ substantiivi liittyy kaikkiin stereotyyppiensä naapureihin ja adjektiivi liittyy kaikkien niiden sanojen naapureihin, joille se on stereotyyppinen ominaisuus.

Naapureiden stereotyypit @? adjektiivi liittyy kaikkiin, joille naapurinsa ovat stereotyyppisiä ominaisuuksia. Substantiivi liittyy kaikkien naapureidensa stereotyyppisiin ominaisuuksiin.

Kategorian naapurit ?^ liittyy kaikkien kategorian sanojen naapureihin.

Naapureiden kategoriat ^? liittyy kaikkiin sanoihin kategorioissa, joihin sen naapurit kuuluvat.

Kategorian stereotyypit @^ liittyy kaikkien kategorian sanojen stereotyyppeihin.

Stereotyyppien kategoriat ^@ liittyy kaikkiin sanoihin kategorioissa, joihin sanan stereotyypit kuuluvat tai joille se on stereotyyppinen ominaisuus.

CIR-menetelmän tehokkuus riippuu hyvin pitkälti siitä, kuinka hyvin edellä esitellyillä operaattoreilla onnistutaan tunnistamaan sanayhteyksiä. Myös @-operaattorissa käytettävä etäisyys on merkittävä, sillä se on ainoa operaattori, joka liittää substantiiveja ja adjektiiveja keskenään.

Strategialla tarkoitetaan korkean tason mallia ironisten vertausten tuottamiseksi. Strategia esitetään CIR-hakuna, jolla etsitään haun operaattoreihin yhteensopivia ilmauksia aineistosta.

Yksi yksinkertaisimmista strategioista uusien ironisten vertausten tuottamiseksi on vastakohtien käyttö. Se voidaan esittää aiemmin esiteltyjen operaattoreiden avulla seuraavasti:

Santonym(P): ?-P @-P

Tässä merkintä Santonym(P) kertoo, että kyseessä on strategia nimeltä antonym (vastakohta) ja annettuna sanana on P. Loppuosasta ilmenee, että ominaisuudesta P johdetaan uusi vertaus hakemalla kaikki Googlen 2-gram aineiston lausekkeet, joissa ensimmäinen sana on adjektiivi, joka liittyy annetun sanan vastakohtaan. Toinen sana on substantiivi, joka esiintyy usein annetun sanan vastakohdan yhteydessä. Näin tällä strategialla voidaan saada johdettua esimerkiksi ilmaus ”soggy pillow” ominaisuudesta ”hard”.

Hieman toisenlaisella strategialla

Scombo(P): @-P @-P

saadaan ominaisuudesta ”soft” luotua monia ironisia ilmauksia kuten ”brick wall”, ”stone wall”, ”steel wall”, ”titanium wall”, ”oak wall” ja ”granite wall”. Sen sijaan ominaisuudelle kova sama strategia tuottaa hieman erikoisempia mielikuvituksellisia ilmauksia kuten ”marshmallow bunny” ja ”snow baby”.

Hieman monimutkaisempi strategia

Sgroup(P) <- (ˆgroup ∩ @P) ”of” @-P

muodostaa kolmen sanan ilmaukset, joissa ensimmäinen sana kuuluu luokkaan ”ryhmä”, esimerkiksi perhe tai armeija, ja jolle annettu ominaisuus on ominainen. Toinen sana ”of” on annettu strategiassa sellaisenaan. Kolmas sana on substantiivi, joka on ominainen annetun ominaisuuden vastakohdalle. Näin syntyy esimerkiksi ilmaukset ”army of dreamers”, ”army of civilians” ja ”army of irregulars”, joille lähtöominaisuutena on ollut ”disciplined”. Sen sijaan ominaisuudelle ”strong” saadaan ’army of cowards”, ”army of babies”, ”army of ants”, ”army of cripples”, ”army of kittens”, ”army of girls” ja ”army of worms”

Strategioihin voidaan oppia automaattisesti erityisiä taktiikoita tutkimalla tiettyjä ironisten kuvausten erityisiä tapauksia. Jos kone pystyy tunnistamaan ironisia vertauksia, se voi myös tunnistaa mitkä ironiset vertaukset hyödyntävät mitäkin strategiaa. Näin voidaan oppia myös sanakohtaisia taktiikoita. Taktiikat toteuttavat korkeamman tason strategiaa, mutta ovat sidottuja tiettyihin sanoihin ja pohjautuvat todennettuihin ironisiin ilmaisuihin. Tässä on hyödynnetty aiemmissa tutkimuksissa kerättyä luokiteltua vertausten aineistoa. Esimerkiksi vertaus, joka kuvaa MS Wordin salasanasuojausta ”about as secure as a cardboard bank vault” (yhtä turvallinen kuin pahvinen pankkiholvi) voidaan nähdä muodossa ”about as secure as a cardboard @secure”, koska sekä ”bank” että ”vault” ovat stereotyyppisesti liittyneet sanaan ”secure”. Näin ”bank vault” tulkitaan yhdistelmäsanaksi, jolla on sama

stereotyyppinen assosiaatio. Koska tämä vertaus oli aineistossa merkitty ironiseksi, sen ironian pitää ilmeisesti syntyä sanan ”cardboard” käytöstä tässä yhteydessä. Sanassa ”cardboard” on siis oltava jotakin, joka heikentää turvallisen turvallisuutta, vaikka siihen ei liitykään ominaisuus turvaton. Näin ollen ”cardboard” osoittaa turvattomuutta vain tässä tietyntyyppisessä yhteydessä kuvatessaan turvallista säiliötä. Tästä johdetaan seuraava ironinen taktiikka

Tcardboard(secure): ”cardboard” @secure

Tämä taktiikka on huomattavasti yksityiskohtaisempi kuin aiemmat strategiat. Siinä liitetään tietty sana (cardboard) ominaisuuteen, jonka pitää olla tiettyä tyyppiä (secure).

Kun strategiaa sovelletaan tiettyyn esimerkkiin, pystyy järjestelmä tunnistamaan esimerkin tarkoituksen ja generoimaan siitä vastaavan taktiikan. Kun taktiikalla Tcardboard(secure) haetaan ilmauksia Googlen 2-grameista, se ehdottaa ironisiksi kuvauksiksi turvallisesta paikasta seuraavia: ”cardboard fortress”, ”cardboard bank”, ”cardboard jail” ja ”cardboard prison”.

Kuitenkin yli 15 tuhannesta ironisesta about-as-vertauksesta vain reilut 10% tuottaa uudelleenkäytettävissä olevan taktiikan, jota voidaan käyttää uusien ironisten vertausten tuottamisessa n-grammeista. Yhteensä taktiikoita saatiin 1694 kappaletta. Jotkin näistä olivat hyvinkin yllättäviä kuten esimerkiksi sanasta ”hard” muodostettava ”foam tombstone” ja sanasta useful saatava ”rubber tripod”.

3 Prototyyppijärjestelmän toteutus

Tässä luvussa esitellään ensin järjestelmän toteutus yleisellä tasolla ja sitten tarkastellaan ja arvioidaan toteutuksen tärkeimpiä vaiheita yksitellen. Järjestelmällä on tarkoitus poimia tekstistä sanojen välisiä yhteyksiä erilaisten mallien perusteella. Saatujen sanojen välisten yhteyksien ja näitä yhteyksiä käyttävien strategioiden avulla tuotetaan sanasta uusi ilmaus, jonka olisi tarkoitus olla ironinen.

3.1 Toteutus

Vealen teknisestä toteutuksesta ja käyttämistä ratkaisusta ei ole varsinaisesti tietoa, koska niitä ei kuvata hänen artikkeleissaan kovinkaan tarkasti, joten toteutuksessa sovelletaan kuvattuja piirteitä oman tulkinnan mukaan. Merkittävimmät erot Vealen

toteutukseen nähden ovat käytettävässä aineistossa. Vealella on käytössään paljon laajempi aineisto, Googlen n-grammit, WordNet ja valmiit varmasti ironiset vertaukset taktiikoita varten. Tässä työssä oli aluksi tarkoitus käyttää aineistona vain suomenkielistä Wikipediaa, mutta myös Project Gutenbergin[1] suomenkielistä aineistoa päätettiin hyödyntää sanayhteyksien vahvistamiseksi (kokeiltiin myös saako kaunokirjallisesta tekstistä poimittua paremmin yhteyksiä kuin tietosanakirjamaisesta tekstistä). Nämä aineistot valittiin niiden laajuuden ja helpon saatavuuden takia. Myöhemmin työn loppupuolella kolmanneksi aineistoksi valikoitui Turku BioNLP Groupin "Finnish Internet Parsebank"-projektin[2] suomenkielinen n-grammiaineisto, joka on kerätty suomalaisilta internet sivuilta. Koska sanayhteyksiä muodostettaessa sanajärjestyksellä on merkitystä, käytettiin aineistosta "flat"-muotoisia 5-grammeja, joissa sanajärjestys säilyy alkuperäisenä. Näissä flat-5-grammeissa on noin 263 miljoonaa riviä (joilla jokaisella on siis 5 sanaa), joista kiinnostavia rivejä on noin 222 miljoonaa. Tässä yhteydessä kiinnostavalla rivillä on ainakin kaksi sanaa, jotka ovat adjektiiveja, substantiiveja tai adverbejä. Tässä vaiheessa ei tosin vielä tarkistettu että adverbit ovat "-sti" päätteisiä eikä sitä että substantiivit eivät ole nimiä tai vierasperäisiä sanoja. N-grammiaineisto ei vaadi esikäsittelyä ja sitä voidaan käyttää sellaisenaan suoraan sanayhteyksien muodostuksessa, toisin kuin Wikipediaa ja Project Gutenberg tekstejä. Myös sanojen tunnistus on jo valmiina n-grammeissa, joten seuraavat esikäsittelyosiot ja sanojen tunnistusosiot eivät koske n-grammiaineistoa.

Aineistoista on tarkoitus saada selvitettyä sanojen välisiä yhteyksiä erilaisten mallien perusteella. Wikipedian vapaasti saatavilla oleva xml-datadump sisältää kaikki suomenkielisen Wikipedian artikkelit ja se käsitellään ensin WikiExtractor[3] skriptillä, joka poistaa ylimääräiset merkinnät (linkit, koodit, viitteet jne.) ja jäljelle jää vain puhdasta tekstiä sisältävä tiedosto. Project Gutenberg tekstit joudutaan valitsemaan ja lataamaan yksitellen manuaalisesti, mutta niitä ei tarvitse siistiä kuten Wikipediaa (käytännöllisyyden vuoksi tekstit yhdistettiin yhdeksi tiedostoksi). Teksteiksi valikoitiin kaunokirjallisia tuotoksia sattumanvaraisesti, runoja ja näytelmiä ei otettu aineistoon.

Wikipedia ja Project Gutenberg teksteistä kuitenkin poistetaan vielä raa'asti kaikki erikoismerkit ja numerot, jotta sanoihin ei jää välimerkeistä ja muista ei-kirjainmerkeistä ylimääräistä "roskaa" häiritsemään sanan tunnistusta. Esimerkiksi sanassa kiinni olevat pilkut, heittomerkit, sulut jne. sanan alussa tai lopussa aiheuttaisivat

[1] http://www.gutenberg.org/ebooks/search/?query=l.finnish
[2] http://bionlp.utu.fi/finnish-internet-parsebank.html
[3] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

ongelmia sanan tunnistamisessa. Ainoastaan sanojen välissä ja lopussa oleva yhdysviiva sallitaan kuten esimerkiksi sanassa linja-auto. Tässä toki menetetään joitain sanojen vivahteita. Esimerkiksi ”5-ovinen” pelkistyy muotoon ”ovinen”. Sanan alussa oleva yhdysmerkki poistetaan, koska tällaisten sanojen perusmuotoja ei onnistuta tunnistamaan. Samalla kuin ei-kirjainmerkit poistetaan, muutetaan kaikki 9 erilaista tekstissä esiintyvää Unicode-merkistön viivaa samanlaisiksi, jotta ne käsitellään yhdenmukaisesti ja oikein sanojen väleissä. Yhtenä esimerkkinä tästä on ajatusviiva ilmaisussa ”Riihimäki–Pietari-rataosuus”, jota ei tunnisteta sellaisenaan, mutta ”perusviivaksi” muutettuna sanat tunnistetaan. Jos ajatusviivaa ei muutettaisi se poistuisi ei-kirjainmerkkien poistamisvaiheessa ja tällöinkin sanat jäisivät tunnistamatta.

Tekstit käydään läpi ja poimitaan jokainen sana/merkkijono, joka tekstissä esiintyy, kaikki erilaiset esiintyvät taivutusmuodot erikseen. Yhden merkin mittaiset merkkijonot hylätään suoraan. Isolla alkukirjaimella olevista sanoista tehdään pienellä kirjoitettu vaihtoehto, koska näillä on havaittu olevan eroa sanojen tunnistamisvaiheessa, esimerkiksi isolla alkukirjaimella kirjoitettuna sanaa napapaju ei tunnistettu, mutta kokonaan pienillä kirjaimilla kirjoitettuna se tunnistettiin oikein. Isolla kirjaimella alkavat sanat ovat usein henkilön tai paikan nimiä, jotka tunnistetaan oikein vain kun ne ovat kirjoitettu isolla alkukirjaimella, joten kaikkia ei voida suoraan vaihtaa pienellä kirjoitetuksi, jos halutaan tunnistaa nimetkin. Isolla kirjaimella alkavat sanat jätettiin tunnistettaviksi ja sanayhteyksissä käytettäviksi, koska haluttiin nähdä saavatko esimerkiksi tunnetut historialliset henkilöt mielenkiintoisia sanayhteyksiä.

Tekstiaineiston läpikäymisen jälkeen saadaan lista tunnetuista sanoista (kokonaisuudessaan siistitty Wikipedia sisältää noin 52 miljoonaa sanaa/merkkijonoa), joista saadaan noin 5 miljoonaa erilaista merkkijonoa, mutta pelkiksi kirjainmerkeiksi siivottuna noin 4,3 miljoonaa erilaista. Project Gutenberg teksteistä saadaan kerättyä noin 46 tuhatta sanaa, jotka eivät esiinny Wikipediassa. Käytetty Project Gutenberg aineisto sisältää yhteensä lähes miljoona sanaa. Edellä olleissa luvuissa on osa sanoista kahteen kertaan, koska isolla alkukirjaimella alkavista sanoista tehtiin pienikirjaiminen versio huolimatta siitä oliko sellainen sana jo aineistossa. Tämä sanalista annetaan jatkokäsittelyyn HFST:lle (Helsinki Finite-State Transducer Technology[4] (HFST)). HFST:llä (hfst-lookup) selvitetään kaikkien sanojen perusmuodot ja sanaluokka, käytännössä tässä kiinnostaa vain se onko sana adjektiivi, substantiivi vai

[4]http://hfst.sourceforge.net/

jotain muuta. Tuloksena voi olla useampia vaihtoehtoisia tulkintoja samalle sanalle, tällöin näistä pitäisi onnistua poimimaan juuri se oikea vaihtoehto. Koska mikään ei varsinaisesti vihjaa sitä mikä olisi paras vaihtoehtoisista tulkinnoista, valittiin aluksi ensimmäinen vaihtoehdoista. Valinta kuitenkin paranee, kun valitaan sellainen sanan perusmuoto, joka esiintyy useimmin tekstissä. Sanojen tunnistusta tarkastellaan tarkemmin omassa luvussaan (luku 3.2).

HFST:n tulosten pohjalta pystytään tunnistamaan adjektiivit ja substantiivit tekstistä ja näiden väliset sanayhteydet voidaan luoda. HFST:n tuloksista saadaan muodostettua tiedostot, joissa on kaikki perusmuotoiset adjektiivit (47 tuhatta), kaikki perusmuotoiset substantiivit (545 tuhatta) ja kaikki tekstissä esiintyvät erimuotoiset tunnistetut sanat ja niiden perusmuodot pareittain (reilut 1,6 miljoonaa kpl).

Sanayhteydet muodostetaan esimerkiksi sen perusteella mitkä sanat esiintyvät peräkkäin aineistona olevassa Wikipediassa. Tässä vaiheessa käytetään sanojen tunnistettuja perusmuotoja, jolloin eri muodoissa olevat samat sanat käsitellään kaikki samana perusmuotoisena sanana. Aluksi sanojen yhteys pääteltiin vierekkäisten sanojen perusteella. Eli kahdella tekstissä peräkkäisellä sanalla, joista kumpikin on joko adjektiivi tai substantiivi, on yhteys keskenään. Näin saadaan paljon sanojen välisiä yhteyksiä, joista osa voi kuitenkin olla hieman yllättäviä. Tämän lisäksi käytettiin sanayhteyksien muodostamisessa myös ”ja”, ”on” ja ”kuin” sanoja. Esimerkiksi ”ja” sana muodostaa yhteyden kahden adjektiivin välille, jos tekstissä esiintyy muoto ”adjektiivi ja adjektiivi”. Sanayhteyksiä tarkastellaan vielä tarkemmin omassa luvussaan (luku 3.3).

Kun sanayhteydet on muodostettu, voidaan niiden pohjalta muodostaa uusia vertauksia käyttämällä strategioita kuten Veale työssään. Operaattorit ovat stereotyyppioperaattori ”@”, naapurioperaattori ”?” ja vastakohtaoperaattori ”-”. Vastakohtia on saatu erillisestä sanalistasta. Se on yhdistelmä kahdesta listasta, joissa on 669 ja 65 sana-vastakohdat paria, yhteensä näissä on 1270 eri sanaa. Nämä listat ovat ”Oppitorin”[5] tuottamia ja Creative Commons -lisenssillä[6] vapaasti käytettävissä.

Esimerkiksi strategia ”@-” sanalla ”hidas” voi tuottaa sanan ’tietoliikenneyhteys’, jolle siis hitaan vastakohta nopea on stereotyyppinen ominaisuus.

Jokaisessa työvaiheessa tuotetaan vaiheen tuloksista tekstitiedosto, jota käytetään seuraavassa vaiheessa työn pohjana, jolloin tästä tiedostosta nähdään mitä on tapahtunut ja mahdolliset virheet on helppo korjata tai voidaan kokeilla jotain tiettyä

[5]http://saaressa.blogspot.fi/2009/02/synonyymit.html
[6]http://creativecommons.org/licenses/by-nc-sa/1.0/fi/

sanaa/asiaa vain lisäämällä se tiedostoon. Tiedostojen käsittely tosin on hieman epäkäytännöllistä tai ainakin aikaa vievää sanojen suuren määrän vuoksi, mutta kriittisimmät ongelmat voidaan oikaista näin (esimerkiksi ongelmallinen sanan väärä tulkinta vaihdetaan tai muokataan sanayhteyttä jne.)

3.2 Sanojen tunnistus

Tässä luvussa esitetään sanojen tunnistamiseen käytetty menetelmä, jolla selvitetään sanan perusmuoto, sanaluokka sekä sijamuoto ja luku(yksikkö/monikko). Ensimmäinen aliluku käsittelee puhdasta tekstiä eli tässä tapauksessa Wikipedia ja Project Gutenberg aineistoja. Toisessa aliluvussa käsitellään n-grammiaineistoa, jolle on jo valmiiksi tehty sanojen tunnistus.

3.2.1 Sanojen tunnistaminen tekstistä

Sanojen tunnistamiseen käytetään siis HFST:tä. Aluksi sanojen tunnistukseen oli tarkoitus käyttää OmorFi[7]-ohjelmaa, mutta ylitsepääsemättömien asennusongelmien vuoksi sitä ei saatu toimimaan käytetyssä ympäristössä ja sanojen tunnistaminen suoritettiin HFST:llä. Suurin vahinko tässä oli se, että Omorfin onnistunut asennusvaihe olisi myös tuottanut ilmeisesti kattavamman morfologiatiedon sanojen tunnistukseen kuin mitä nyt oli mahdollista käyttää. Näin sanojen tunnistaminen ei onnistu ilmeisestikään aivan parhaalla mahdollisella tavalla. Seuraavassa esimerkki saaduista tuloksista, kun HFST-lookup:lle on annettu sanat Napapaju ja napapaju analysoitavaksi.

```
Napapaju Napapaju+? inf

napapaju napa#paju N Nom Sg 312.000000
```

Isolla alkukirjaimella olevaa sanaa ei siis ole tunnistettu ja pienellä kirjoitettaessa sana on tulkittu substantiiviksi, jonka perusmuoto on kaksiosainen sana napapaju. Merkinnät Nom ja Sg kertovat lisäksi sanan olevan nominatiivi ja yksikössä. Lopussa oleva luku on kaikille tunnistetuille sanoille sama, eikä näin ollen kerro mitään oleellista sanasta. Koska isot alkukirjaimet häiritsevät tunnistusta, on isolla kirjaimella alkavista sanoista jouduttu tekemään myös kokonaan pienellä kirjoitetut versiot tunnistusta varten. Sana napapaju ei edes esiinny sellaisenaan pienellä

[7] http://code.google.com/p/omorfi/

kirjoitettuna aineistossa kertaakaan, ainoastaan muodossa napapajua, joka tunnistetaan oikein. Kaiken kaikkiaan erilaisia sanoja on noin 4,3 miljoonaa ja näistä noin 2,3 miljoonaa jää tunnistamatta. Tunnistamattomissa sanoissa on mukana paljon vierasperäisiä nimiä, joitain isolla alkukirjaimella alkavia sanoja, joitakin kirjoitusvirheitä, osa sanoista on epämääräisiä merkkijonoja ja joitakin ”normaaleja” sanoja, joita ei jostain syystä onnistuta tunnistamaan. Nimien kohdalla on usein pienillä kirjaimilla kirjoitettu vaihtoehto tunnistamatta, kun isolla alkukirjaimella oleva on tunnistettu oikein nimeksi. Tunnistamattomia sanoja ovat esimerkiksi Lutsenkoa, lutsenkoa, Maalaiskunnantie (pienellä kirjoitettuna tunnistetaan), kitetyä, kansainvälisisssä, U-B ja monisyisestikin. Myös sanat jediakatemia, stemmausvirheitä ja ”kjik-kjik-kjik” (niittysuohaukan ääntely) jäävät tunnistamatta. Näiden lisäksi Project Gutenberg -aineistossa on vanhahtavaa tai murteellista tekstiä, jota HFST:llä ei tunnisteta, kuten esimerkiksi sanat ”tarvitte”, ”lauvennut” ja ”aprikoita”.

Sanojen, joiden alussa on lyhenne, ehdotetut tulkinnat ovat hieman erinäköisiä kuin muiden ja jotkut merkkijonot voidaan tulkita lyhenteeksi, esimerkiksi:

```
NHL-otteluaan NHL Abbr Prop TrunCoottelu N Par Sg Px3 312.000000

zC zC Abbr 312.000000
zC zC Abbr Nom Sg 312.000000

EU-tyyppihyväksyntä EU Abbr Prop TrunCotyyppihyväksyntä N Nom Sg 312.000000
EU-tyyppihyväksyntä EU Abbr TrunCotyyppihyväksyntä N Nom Sg 312.000000
```

Myöhempiä vaiheita varten edellä esitetyn kaltaiset tulosten muodot yhdenmukaistetaan muiden sanojen kanssa niin, että ”Abbr [Prop] TrunCo”-osa vaihdetaan pelkkään ”-”-merkkiin. Eli sanan ”EU-tyyppihyväksyntä” tulos on ”EU-tyyppihyväksyntä N Nom Sg 312.000000”. Edellisissä esimerkeissä ”Abbr” kertoo lyhenteestä, ”Prop” ilmaisee että sana ”EU” on ainutlaatuinen kokonaisuus eikä jotakin ryhmää kuvaava substantiivi. Eli ”Prop” kertoo kyseisen sanan olevan erisnimi. Esimerkiksi ”Jupiter” olisi tällainen ainutlaatuinen yksilöivä sana, kun taas ”planeetta” olisi tavallinen luokan jäseniä kuvaava sana. Näyttää siltä, että sanan iso alkukirjain vaikuttaa lähes aina siten, että sana tulkitaan erisnimeksi. Loppusanaan yhdistetty ”TrunCo” kuvaa katkaistua yhdistettä (truncated compound). HFST:n tuottamien sanojen ominaisuuksien lyhenteiden merkityksien tulkinnassa auttoi ”FinnTreeBank2 Manual” [VPM12].

Myös e-hydroksidimetyyliallyylipyrofosfaatti tunnistetaan onnistuneesti ja oikein, tosin sen oikea kirjoitusasu olisi (E)-4-hydroksi-3-metyylibut-2-enyylipyrofosfaatti eli (E)-4-hydroksidimetyyliallyylipyrofosfaatti, mutta siistimisen seurauksena aineistossa esiintyy pelkistetty muoto, joka on myös ainoa onnistuneesti tunnistettu muoto edellä mainituista.

Tunnistettuja sanoja on reilut 2 miljoonaa ja näille saadaan keskimäärin 2,6 vastausvaihtoehtoa sanaa kohden. Esimerkiksi sana ”sotilasajoneuvoista” tuottaa seuraavat vastausvaihtoehdot:

```
sotilasajoneuvoista sotilas#ajo#-neuvoinen A Pos Par Sg 312.000000
sotilasajoneuvoista sotilas#ajo#neuvo N Ela Pl 312.000000
sotilasajoneuvoista sotilas#ajo#neuvo N Ela Pl 312.000000
sotilasajoneuvoista sotilas-#ajo#-neuvoinen A Pos Par Sg 312.000000
sotilasajoneuvoista sotilas-#ajo#neuvo N Ela Pl 312.000000
```

Osa vaihtoehdoista on tunnistettu adjektiiviksi ja osa substantiiviksi, myös ehdotetuissa perusmuodoissa on hieman eroja. Oikea tulkinta olisi mitä ilmeisimmin substantiivi ”sotilasajoneuvo”, joka on kahdessa annetuista vaihtoehdoista. Usein ei kuitenkaan ole niin, että oikea vaihtoehto olisi enemmistönä vaihtoehdoissa, kuten esimerkiksi sanan ”kilttinä” tuloksissa, joissa vaihtoehtoja on kaksi. Molemmissa perusmuotona on kiltti, mutta toinen on adjektiivi ja toinen on substantiivi. Joillekin sanoille saadaan useita vaihtoehtoja, jotka kaikki kuitenkin ovat täysin samoja, esimerkiksi sana ”juniorijääkiekkojoukkueen” saa neljä vaihtoehtoa, kaikki substantiiveja joiden perusmuoto on ”juniorijääkiekkojoukkue”.

Kaiken kaikkiaan tunnistettuja sanoja on Wikipediassa lähes kaksi miljoonaa. Jos kaikki erilaiset tulkintavaihtoehdot huomioitaisiin, näistä saataisiin yli kolme miljoonaa eri tulkintaa, samasta sanasta olisi siis useampi vaihtoehtoinen tulkinta. Jos käytetään yksinkertaista tapaa ja valitaan aina tulosten ensimmäinen vaihtoehto, saadaan yli 560 tuhatta substantiivia ja 49885 adjektiivia. Kun vielä adjektiiveista kaikki muutetaan pieniksi kirjaimiksi, ettei samasta sanasta ole kahta vaihtoehtoa, saadaan adjektiivien määräksi 48995. Erilaisia substantiiveiksi tai adjektiiveiksi tunnistettuja sanojen taivutusmuotoja on tällöin yhteensä yli 1,6 miljoonaa kappaletta ja perusmuotoja on reilut 611 tuhatta. Noin 80% kaikista tunnistetuista sanoista on substantiiveja tai adjektiiveja.

Project Gutenberg -aineisto lisää sanojen kokonaismäärää noin 46 tuhannella ja näistä tunnistetaan HFST:llä noin 26 tuhatta.

Koska mikään ei näytä vihjaavan tulosten parhaasta vaihtoehdosta, päädyttiin aluksi valitsemaan aina ensimmäinen vaihtoehto, joka on kuitenkin usein oikein. Tällöin joistakin sanoista kuten esimerkiksi ”kuin” aiheutui suuret ongelmat kun HFST tuloksissa ensimmäinen vaihtoehto on sana ”kui” (substantiivi) ja kaikenlisäksi sanan ”kuu” taivutusmuodoissa ensimmäinen tulosvaihtoehto on myös sana ”kui”, jostain syystä myös ”kuu” päätteiset sanat saavat vaihtoehdoksi ”kui” päätteen. Näin sana ”kui” yhdistyi sanayhteyksiä tehdessä todella moniin sanoihin. Seuraavassa joitakin esimerkkejä sanoista, joissa ”kui”:

```
Kuin Kuu N Prop Ins Pl 312.000000

kuin kui N Gen Sg 312.000000
kuin kui N Ins Pl 312.000000
kuin kuin CS 312.000000
kuin kuu N Ins Pl 312.000000

kuiden kui N Gen Pl 312.000000
kuiden kuu N Gen Pl 312.000000

kostoiskuissa kos#toinen#kui N Ine Sg 312.000000
kostoiskuissa kos#toinen#kui N Ine Pl 312.000000
kostoiskuissa kos#toinen#kuu N Ine Pl 312.000000
kostoiskuissa kosto#isku N Ine Pl 312.000000

ikkunaluukuilla ikkuna#luu#kui N Ade Sg 312.000000
ikkunaluukuilla ikkuna#luu#kui N Ade Pl 312.000000
ikkunaluukuilla ikkuna#luu#kuu N Ade Pl 312.000000
ikkunaluukuilla ikkuna#luukku N Ade Pl 312.000000
ikkunaluukuilla ikkuna#luukku N Ade Pl 312.000000

helmikuista helmi#-kuinen A Pos Par Sg 312.000000
helmikuista helmi#kui N Ela Sg 312.000000
helmikuista helmi#kui N Ela Pl 312.000000
helmikuista helmi#kuu N Ela Pl 312.000000
helmikuista helmi#kuinen A Pos Par Sg 312.000000
helmikuista helmi#kuu N Ela Pl 312.000000
```

Kuten edellä esitetyistä esimerkeistä huomataan, sanalle oikean tulkintavaihtoehdon valitseminen on monissa tapauksissa hyvin hankalaa, vaikka tuloksissa onkin mukana juuri se oikea vaihtoehto. Siksi sanojen tunnistus näyttää olevan yksi tämän toteutuksen heikko kohta, varsinkin kun katsotaan miten sanoja tulkitaan välillä täysin väärin ja osaa sanoista ei ymmärretä ollenkaan. Erityisesti valinta tuloksista, joissa on monia täysin erilaisia vaihtoehtoja (substantiivi/verbi/adjektiivi) on hankalaa ja näin voi jäädä joku haluttu sana tunnistamatta oikein substantiiviksi tai adjektiiviksi. Se että sanan perusmuoto on hieman väärin, ei ole kuitenkaan lopulta kovinkaan suuri ongelma, jos se on kuitenkin helposti ihmisen pääteltävissä. Sanojen tunnistus paranisi ilmeisesti paljon, jos käytössä olisi parempi morfologia-tiedosto. Toisaalta pikaisella yleissilmäilyllä sanojen tulkinnoista näyttää kuitenkin siltä, että kokonaiskuvassa sanojen tunnistus toimii melko hyvin. Lopullisista tuloksista, eli ironisista vertauksista, pitää ihmisen mahdollisesti tulkita hieman, koska jotkin sanat voivat näyttää kummallisilta.

Kun sanojen perusmuotojen tulkinnassa huomioidaan se esiintyykö perusmuoto sellaisenaan tekstissä, saadaan osa oudoimmista tulkinnoista karsittua. Esimerkiksi ”heinäkui” vaihtoehto jätetään valitsematta ”heinäkuu”:n sijasta. Kuitenkin ”kuin” sana saa perusmuodokseen edelleen sanan ”kui”, koska ”kui” esiintyy tekstissä sellaisenaan. Tämä vältetään, kun huomioidaan myös sanojen esiintymismäärät tekstissä ja valitaan sanalle se perusmuoto, joka esiintyy useimmiten tekstissä sellaisenaan.

Ongelmana on kuitenkin edelleen yhdyssanojen muodot, joissa kaikki sanan osat on muutettu perusmuotoon HFST:n toimesta. Esimerkiksi sana ”valtakunnanasuntokomissaariksi”, josta tulee kolme vaihtoehtoista muotoa, joista kaksi on ”valta#kunta#asunto#komissaari” ja yksi ”valta-#kunta#asunto#komissaari”. Myös jotkin isolla kirjaimella alkavat tai väliviivan sisältävät sanat saavat erikoisen näköisiä tulkintoja esimerkiksi ”Keski-Lännessä” on perusmuodossa ”Keski#Länne” tai ”Keski#Länsi”, mutta ”keski-lännessä” on ”keski-#länsi”. Samoin ”Sievissä” tulkitaan perusmuodoksi ”Sievi”, mutta pienellä kirjoitettaessa ”sievissä” saa muodon ”sievä”.

Joillekin sanoille saadaan vain yksi tulkinta, joka on väärä. Esimerkiksi sanan ”pienveturi” ja muiden sen taivutusmuotojen tulkitaan olevan taivutettu muoto yhdyssanasta ”piki#veturi”, vaikka ”pienveturi” olisi oikea muoto.

Sanan esiintymismäärien huomioiminen oikean perusmuodon valitsemisessa pienentää kaikkien erilaisten perusmuotojen kokonaismäärää. Saman sanan eri taivutusmuodot eivät näin saa yhtä herkästi eri perusmuotoja, kuin vain valitsemalla ensimmäinen vaihtoehto. Esimerkiksi pelkästään Wikipediasta saadaan kerättyä yli 562

tuhatta substantiivia ja noin 49 tuhatta adjektiivia perusmuodoissa (kun valitaan ensimmäinen vaihtoehto perusmuodoksi). Mutta huomioitaessa perusmuodon esiintymismäärät Wikipedia ja Project Gutenberg tiedostot yhdessä tuottavat vähemmän eri perusmuotoja (545 tuhatta substantiivia ja 45 tuhatta adjektiivia), vaikka erilaisten sanojen kokonaismäärä on kasvanut. Myös sanojen määrä eri taivutusmuodoissa on kasvanut, mutta erilaisten valittujen tulkintojen määrä siis vähentynyt.

Sanojen esiintymismäärien lisäksi oikeaa muotoa valittaessa huomioidaan HFST perusmuotoehdotuksissa olevien #-merkkien määrä. Sillä aineistoa tutkimalla huomattiin, että vähiten #-merkkejä sisältävä tulkinta on usein paras. Tämä ei vaikuta juurikaan niissä tapauksissa, joissa sanan perusmuoto löytyy aineistosta. Tällöin vain esitysmuoto on hieman selkeämpi, vaikka lopputulos sanasta on täysin sama. Ne sanat, joille mikään vaihtoehtoisista perusmuototulkinnoista ei löydy aineistosta saavat tällä tavoin usein parhaimmalta vaikuttavan tuloksen. Useampia #-merkkejä sisältävä tulkinta on usein ylianalysoinut sanaa liian monista eri sanoista koostuvaksi sanaksi.

Esimerkiksi sana ”matala” saa kaksi vaihtoehtoista tulkintaa ”mata#la N” ja ”matala A Pos Nom Sg”. Tulkinnasta riippuen kyseessä on siis joko substantiivi tai adjektiivi, esiintymismäärät molemmille perusmuotovaihtoehdoille ovat luonnollisesti samat. Ennen #-määrien huomioimista (jolloin valittiin ensimmäinen vaihtoehto) ”matala” kuitenkin löytyi perusmuotoisten adjektiivien joukosta, koska muiden muassa sanasta ”matalammissa” oli tulkittu perusmuoto ”matala A”. Mutta tekstissä olevan sanan ”matala” oli tulkittu olevan substantiivi. Näin ollen perusmuotoon muutettuna sana ”matala” oli siis sekä substantiivi että adjektiivi, mutta esiintyessään tekstissä muodossa ”matala” sitä ei tulkittu adjektiiviksi.

Muita sanoja, jotka saadaan näin tulkittua oikein, ovat esimerkiksi sana ”kielitesti”, jolle tulkitaan perusmuodoksi ”kieli#tes#ti N”. Mutta kun valitaan muoto, joka muodostuu pienimmästä määrästä sanoja valitaan perusmuodoksi ”kieli#testi N Nom Sg”. Toinen tällainen sana on ”määrätietoisena”, josta saadaan ”määrä#tietoinen A Pos Ess Sg” mutta myös ”määrä#tie#toinen N Ess Sg” olisi vaihtoehtoinen tulkinta, joka kuitenkin nyt hylätään. Myös yksi tällainen sana on ”viestitietokannasta”, jolle alun pitäen saatiin hieman omituiselta vaikuttava tulkinta ”viestitietokanasta”, koska mitään perusmuotovaihtoehtoa ei löytynyt sellaisenaan tekstiaineistosta.

```
viestitietokannasta viesti#tie#toka#nasta A Pos Nom Sg 312.000000
viestitietokannasta viesti#tie#toka#nasta N Nom Sg 312.000000
```

```
viestitietokannasta viesti#tie#tokka#nasta A Pos Nom Sg 312.000000
viestitietokannasta viesti#tie#tokka#nasta N Nom Sg 312.000000
viestitietokannasta viesti#tieto#kanna N Ela Sg 312.000000
viestitietokannasta viesti#tieto#kannas N Par Sg 312.000000
viestitietokannasta viesti#tieto#kanta N Ela Sg 312.000000
viestitietokannasta viesti#tieto#kanta N Ela Sg 312.000000
```

Tässä jäljelle jäävistä vähiten #-merkkejä sisältävistä vaihtoehdoista voidaan vielä valita se muoto, joka esiintyy useimmin. Näin saadaan valittua perusmuoto ”viestitietokanta”.

Tämä #-merkkien määrien huomioiminen vaikutti noin 32 tuhanteen sanaan, mutta useimpien näistä kohdalla tulkinta ei kuitenkaan muuttunut merkittävästi. Seuraavassa joitakin esimerkkejä näistä sanoista:

```
lentovarikko:
 lento#vari#kko N Abbr Nom Sg -> lento#varikko N Nom Sg
doping-testi:
 doping#tes#ti N -> doping#testi N Nom Sg
jääpallomestariksi:
 jää#pallo#mesta#riksi N Nom Sg -> jää#pallo#mestari N Tra Sg
kuljetuspanssariajoneuvot:
 kuljetus#panssari#ajo#neuvo N Nom Pl
                -> kuljetus#panssari#ajoneuvo N Nom Pl
tukkilaisena:
 tukki#lainen A Pos Ess Sg -> tukkilainen N Ess Sg
karikkoineen:
 kari#kko N Abbr Com Px3 -> karikko N Com Px3
```

Edellisessä esimerkissä olevan sanan ”jääpallomestariksi” perusmuotoa ei saatu esiintymismääriä laskemalla oikeaksi, koska molemmat vaihtoehdot esiintyivät aineistossa vain yhden kerran. Suurimmalle osalle muuttuvista sanoista muutos ei kuitenkaan näytä vaikuttavan oleellisesti tulkintaan. Joissain tapauksissa on kuitenkin useampi erilainen vaihtoehto, jotka saavat yhtä monta esiintymää. Tällöin voi väärä vaihtoehto tulla valituksi.

Myös yhdysmerkkien (”-”) määrä huomioidaan perusmuotovaihtoehtoa valittaessa, koska HFST:llä tulee useisiin sanojen perusmuotoihin ylimääräisiä yhdysmerkkejä. Valitaan vähiten yhdysmerkkejä sisältävä sana, jos vaihtoehtoja on vielä muiden

sääntöjen (perusmuodon esiintymismäärä tekstissä ja #-määrä) läpikäymisen jälkeen jäljellä.

Jos vielä näiden valintojen jälkeen jäljellä on useampia vaihtoehtoja, valitaan sellainen muoto, joka esiintyy useimmin tulkintavaihtoehdoissa. Kuten edellä olleessa esimerkissä, jossa sana ”viestitietokannasta” sai kaksi täysin samaa tulkintaa. Jos kaikkia vaihtoehtoja on vain yksi kappale, valitaan viimeinen vaihtoehto. Joidenkin sanojen tulkinta muuttuu oikeaksi valittaessa useimmin esiintyvä muoto, joillekin taas olisi ollut parempi valita viimeinen vaihtoehto jäljellä olevista. Tämä useimmin esiintyvän vaihtoehdon valinta vaikuttaa noin 7 tuhanteen sanaan. Monissa näistä muutoksista on kyse yksikkö-monikko muutoksesta tai nominatiivin ja genetiivin vaihtumisesta keskenään.

Seuraavassa joitakin esimerkkejä muutoksista, joissa on muodon esiintymismäärä huomioituna:

```
oppijärjestelmänsä:
 oppi#järjestelmä N Nom Sg Px3 (2kpl tätä vaihtoehtoa)
    -> oppi#järjestelmä N Nom Pl Px3 (2kpl)

maailmanperintöehdokkaana:
 maa#ilman#perintö#ehdokas N Ess Sg (1kpl)
    -> maa#ilma#perintö#ehdokas N Ess Sg (2kpl)
     (myös maa#ilma#perintö-#ehdokas 2kpl)

lisäraitoina:
 lisä#raito N Ess Pl (1kpl)-> lisä#raita N Ess Pl (2kpl)

puutavaraosakeyhtiön:
 puuta#vara#osakeyhtiö N Gen Sg (1kpl)
    -> puu#tavara#osakeyhtiö N Gen Sg (2kpl)
```

Adjektiivien määrä nousee vielä hieman (noin 1800 sanaa), kun ”sti” ja ”stikin” päätteiset adverbit käsitellään kuten adjektiivit. Näin saadaan esimerkiksi sana ”nopeasti” mukaan yhteyksiin. HFST tunnistaa sanan ”nopeasti” perusmuodoksi sanan ”nopeasti”. Sanoille ”nopea” ja ”nopeasti”, jotka käsitellään erillisinä, saadaan lähes yhtä paljon yhteyksiä.

Lopullisena tuloksena saadaan 550133 perusmuotoista substantiivia ja 42136 perusmuotoista adjektiivia tai adverbia. Erilaisissa taivutusmuodoissa olevia substantiiveja, adjektiiveja ja adverbeja on tunnistettu yhteensä 1632604 kappaletta.

3.2.2 Sanojen tunnistaminen n-grammiaineistosta

N-grammiaineisto on valmiiksi luokiteltuna, joten sille ei tarvitse tehdä sanojen tunnistusta. Tosin jotkin sanojen tulkinnat ovat selkeästi väärin ja edellä kuvattua menetelmää käyttäen osa vääristä tulkinnoista olisi saatu tulkittua oikein. Tällöin on kuitenkin mahdollista, että jotkin muut sanat olisi tulkittu väärin.

N-grammiaineisto koostuu seuraavanlaisista riveistä, jotka sisältävät viisi ”sanaa”. (Tässä yksi rivi on pilkottu viidelle riville selkeyden vuoksi)

```
edistyneet/edistyä/V/NUM_Pl|CASE_Nom|VOICE_Act|PCP_PrfPrc|CMP_Pos
ilmansuodattimet/ilman|suodatin/N/NUM_Pl|CASE_Nom
kypärän/kypärä/N/NUM_Sg|CASE_Gen
t-muotoisessa/t|muotoinen/A/NUM_Sg|CASE_Ine|CMP_Pos|CASECHANGE_Up
visiirissä/visiiri/N/NUM_Sg|CASE_Ine
```

Rivillä on siis viisi kertaa muoto ”sana / sananperusmuoto / sanaluokka / luku | sanamuototietoja”. Rivien lopussa on myös luku, joka kertoo kunkin rivin esiintymismäärän. Sanaan liittyvistä tiedoista käytetään sanaluokkaa tunnistamaan onko kyseessä adjektiivi(”A”) tai substantiivi(”N”). Sanan sijamuotoa(”CASE_xxx”) ja lukua(”NUM_xx”) käytetään tarkistettaessa voiko kahden sanan välille luoda sanayhteyden.

Verrattuna edellisessä aliluvussa esiteltyyn HFST:tä käyttävään menetelmään näyttää siltä, että nämä n-grammien valmiit sanojen luokittelut ovat onnistuneempia erityisesti nimien kohdalla. Esimerkiksi sana ”mp2” on tunnistettu nimeksi toisin kuin HFST:llä. Myös numeroita sisältävät sanat on tunnistettu paremmin kuin mihin HFST pystyi. Esimerkiksi sana ”14-16-vuotiaat” on tunnistettu muutoin oikein, mutta perusmuodoksi tulee yhdysviivat hävittävä sana ”1416vuotias”. Näin myös sanan merkitys muuttuu huomattavasti. Aineistossa on myös joitakin virheellisiä tulkintoja, jotka havaittiin vain kevyellä aineiston silmäilyllä. Esimerkiksi sana ”joista” saa perusmuodoksi sanan ”joki”, vaikka edeltävänä sanana olisikin pilkku. Muita väärin tulkittuja sanoja, jotka olisi edellisessä aliluvussa kuvatulla HFST-menetelmällä saatu tunnistettua oikein, ovat esimerkiksi sanat ”henkisenkin”, ”kan-

sainvälisenkin", "pannupizzasiivu" ja "astrologiseksi". Näille tulkitut perusmuodot ovat "henki|senkki", "kansa|väli|senkki", "panna|pizza|siivu" ja "astrologi|seksi".

Sen sijaan sanat "hoitamisessa" ja "hoitamattomassa" saavat molemmat perusmuodoksi sanan "hoitaa", mutta ensimmäinen on substantiivi ja jälkimmäinen on adjektiivi. HFST-menetelmällä olisi sanat voitu tulkita samoin substantiiviksi("hoitaminen") ja adjektiiviksi("hoitamaton"), mutta perusmuodot olisivat olleet erilaiset, tai vaihtoehtoisesti perusmuoto olisi ollut sama "hoitaa" sana, mutta se olisi tulkittu verbiksi.

3.2.3 Valittu tunnistusmenetelmä

Wikipedialle ja Project Gutenberg -teksteille, jos sanalle saadaan useampia vaihtoehtoisia muotoja HFST:llä, käytetään seuraavaa viisivaiheista menetelmää:

Ensiksi katsotaan sanalle saatujen perusmuotovaihtoehtojen esiintymismäärä teksteissä ja valitaan eniten esiintyvä muoto.

Jos jäljellä on vielä useampi vaihtoehto, valitaan vähiten "#"-merkkejä sisältävä vaihtoehto.

Jos jäljellä on vielä useampi vaihtoehto, valitaan vähiten "-"-merkkejä sisältävä vaihtoehto.

Jos jäljellä on vielä useampi vaihtoehto, valitaan useimmin näissä vaihtoehdoissa esiintyvä muoto.

Jos jäljellä on vielä useampia vaihtoehtoja, tulee viimeisin vaihtoehto valituksi.

N-grammiaineistolle ei tehdä sanojen tunnistusta, sillä se on jo valmiiksi käsiteltynä ja näitä valmiita tuloksia hyödynnetään sellaisenaan.

3.3 Sanayhteydet

3.3.1 Sanayhteydet Wikipediasta ja Project Gutenberg -teksteistä

Sanayhteyksien muodostamiseen käytettiin aluksi järjestelmän toimintaa testattaessa yksinkertaista mallia, jossa tekstissä vierekkäin esiintyvien sanojen tulkitaan sisältävän yhteys. Näin saadaan kerättyä ainakin monia yhteyksiä, joista tosin useat

voivat olla hieman epäilyttäviä. Näin kerättynä saadaan yhteyksiä yli 400 tuhannelle sanalle. Yhteensä sanoja on yhteyksissä yli 8,1 miljoonaa siis keskimäärin 20 yhteyttä joka sanalle. Näissä on kuitenkin paljon sanakohtaisia eroja. Esimerkiksi substantiivi lähi-ilmatorjuntaohjus yhdistyy vain adjektiiviksi tulkittuun sanaan ”oleva”, tämä yhteys on saatu tekstin osasta ”käytettävissä olevista lähi-ilmatorjuntaohjuksista laukaistaan”. Sen sijaan useammasta sanasta väärin tulkittu perusmuoto ”kui” sai kolmetuhatta yhteyttä adjektiiveihin ja lähes 17 tuhatta yhteyttä substantiiveihin, ennen kuin sanojen perusmuotojen tulkintaa korjattiin huomioimaan sanojen esiintymismäärät.

Myös sanayhteyksissä ilmenee ongelmia pienten ja isojen alkukirjaimien kanssa. Esimerkiksi ”IT-johtaja” yhdistyy vain sanaan ”valtio”, samoin ”it-johtaja”. Sen sijaan sana pöllö on sanan älykäs naapuri, mutta Pöllö on sille stereotyyppinen. Sanoilla pöllö ja Pöllö on siis tulkittu olevan eri tarkoitus, toinen on substantiivi ja toinen adjektiivi, adjektiivilla pöllö on myös paljon enemmän sanayhteyksiä. Näiden yhteyksistä ilmenee, että pöllöryhmät ovat stereotyyppisesti pöllöjä, mutta Pöllöt ovat älykkäitä. Tämä pöllöjen ominaisuus katoaa, kun sana ”pöllö” ja sen taivutusmuodot tulkitaan substantiiviksi eikä adjektiiviksi.

Yhteyksiä tarkastelemalla kuitenkin huomataan, että tämä vierekkäisten sanojen käyttö yhteyksien muodostamisessa ei yksinään ole riittävän hyvä menetelmä ja on syytä käyttää muitakin tapoja. Mahdollisesti myös yksittäiset osumat on syytä poistaa oudoimpien yhteyksien vähentämiseksi.

Veale käytti seuraavia malleja sanayhteyksien keräämisessä [Vea12]:

adjektiivi-substantiivi yhteys ”DET PROPERTY NOUN”

adjektiivien välinen yhteys ”as X and Y as”

substantiivien välinen yhteys ”Xs and Ys” , jossa X ja Y ovat substantiiveja monikossa

Tässä työssä sanayhteyksien muodostuksessa kokeiltiin seuraavia erilaisia vaihtoehtoisia malleja:

”substantiivi substantiivi” vierekkäiset sanat

”adjektiivi substantiivi”

”ajektiivi adjektiivi”

”substantiivi(Gen) substantiivi(Nom)” sanojen sijamuodon oltava genetiivi ja nominatiivi.

”substantiivi ja substantiivi”

”substantiivi(Pl) ja substantiivi(Pl)” molemmat monikossa.

”adjektiivi ja adjektiivi”

”adjektiivi kuin substantiivi”

”adjektiivi(Comp) kuin substantiivi” adjektiivin oltava komparatiivi.

”yhtä adjektiivi kuin substantiivi”

”substantiivi on/oli/ovat/olivat adjektiivi”

”substantiivi on/oli/ovat/olivat substantiivi”

”substantiivi tai substantiivi”

”adjektiivi tai adjektiivi”

”yhtä adjektiivi ja adjektiivi kuin”

”adjektiivi mutta adjektiivi”

”adjektiivi vaikka adjektiivi”

”adjektiivi(Superl) adjektiivi(Superl)” sanojen oltava superlatiiveja.

Näitä käyttämällä saataisiin poimittua esimerkiksi yhteydet ”kova kuin kivi”, ”yhtä nopeasti ja voimakkaasti kuin” ja ”lääkärit ja hoitajat” tapaisista tekstin osista. Edellisistä ainoastaan kovan ja kiven yhteys on saatu vierekkäisistä sanoista. Lääkäri yhdistyy taloudenhoitajaan muttei (sairaan)hoitajaan, kun vain viereiset sanat ovat käytössä yhteyksiä muodostettaessa.

Yhteyksiä muodostettaessa vaaditaan, että sanat ovat samassa taivutusmuodossa. Tämä vähentää erilaisten sanojen määrää yhteyksissä huomattavasti ja näyttää parantavan yhteyksien laatua (yhteydet vaikuttavat loogisilta eivätkä sattumanvaraisilta). Sanojen ja yhteyksien suuresta määrästä johtuen eri yhteyksienmuodostustapojen arvioinnissa ei ole käyty kaikkia tuloksia läpi, vaan kokonaiskuva on muodostettu satunnaisesti tutkittaviksi valittujen yhteyksien pohjalta. Yhteyksien laadun

arvioinnissa ei ole käytetty mitään tarkkaa menetelmää, vaan arviointi perustuu kirjoittajalle sanoista syntyviin mielikuviin.

Kuten todettua **vierekkäisistä sanoista** muodostettavat yhteydet eivät toimi erityisen hyvin, sillä mukaan tulee hyvin paljon erikoisia yhteyksiä. Esimerkiksi ”tilamonikäyttöauto” yhdistyy sanaan ”pieni” mikä ei vaikuta erityisen kuvaavalta tila-autolle. Sen sijaan ”pikkuauto” yhdistyy sanoihin ”edullinen”, ”halpa” ja ”takamoottorinen”. Vierekkäisten sanojen yhteyksillä saadaan kerättyä paljon yhteyksiä ja erilaisia sanoja. Sanapari ”adjektiivi substantiivi” tuottaa miltei 2 miljoonaa osumaa, ”adjektiivi adjektiivi” yli 400 tuhatta ja ”substantiivi substantiivi” 1,7 miljoonaa osumaa, vaikka sanojen vaadittiin olevan samassa taivutusmuodossa. Kun ”substantiivi substantiivi” parin vaaditaan olevan genetiivi ja nominatiivi putoaa osumamäärä noin puoleen aiemmasta.

Käyttämällä vierekkäisten substantiivien yhdistämisessä vaatimusta, että ensimmäisen sanan oli oltava genetiivi ja toisen sanan nominatiivi, haluttiin saada yhteyksiä kuten ”auton kuljettaja” ja ”auton ratti”. Näin saatiinkin monia hyviä yhteyksiä, mutta monet yhteydet olivat edelleen myös huonoja. Esimerkiksi ”pikkuauto” yhdistyy nyt substantiiveihin ”kehitystyö”, ”suunnittelu”, ”tila”, ”moottori”, ”prototyyppi” ja ”ensimmäinen”.

Käytettäessä **”ja”** sanaa yhdistämään kaksi substantiivia(**”substantiivi ja substantiivi”**) tai kaksi adjektiivia(**”adjektiivi ja adjektiivi”**) toisiinsa, saadaan Wikipediasta poimittua kaiken kaikkiaan miltei 60 tuhatta adjektiivi paria ja yli 400 tuhatta substantiivi paria. Edellisissä luvuissa saattoi esiintyä samat sanaparit useamman kerran. Näin suuriin määriin mahtuu mukaan paljon huonojakin yhteyksiä. Monet ”ja” sanan avulla saadut yhteydet ovat kuitenkin lupaavia. Näin saadaan esimerkiksi ”lääkäri-hoitaja” -yhteys ja myös ”valelääkäri-identiteettivarkaus” -yhteys. Samoin esimerkiksi sana ”käytettävyys” saa 45 yhteysosumaa, joista vahvimmat yhteydet ovat sanoihin ”saatavuus”, ”toiminnallisuus”, ”ulkonäkö”, ”käyttöliittymä” ja ”ergonomia”. Mallin ”substantiivi ja substantiivi” yhteyksissä on monia huonojakin yhteyksiä, mutta kun ehtona on, että molemmat sanat ovat monikossa, on joidenkin sanojen yhteyksissä parhaimmat jäljellä. Näin ei tosin ole kaikkien sanojen kohdalla. Monikkomuoto näyttää silti olevan suhteellisen hyvä tapa kerätä sanayhteyksiä, sillä muillakin tavoilla tulee mukaan huonoja yhteyksiä. Kun sanojen on oltava monikossa sana ”luostari” saa yhteydet ”kirkko” (3kpl), ”maanomistaja” ja ”aateliskartano”. Samoin sana ”tiivistelmä” yhdistyy vain sanaan ”otsikko”. Ilman monikkomuodon vaatimusta ”luostari” saa edellisten lisäksi yhteydet sanoihin ”kirkkokunta”, ”avio-

liitto”, ”hallitsijapari” ja ”kappeli”. Tällöin ”tiivistelmä” saa myös yhteydet sanoihin ”nimi”, ”säädöskäännös”, ”kirjallisuustiedo”, ”perusteos”, ”tunnustus”, ”asiasanaluettelo”, ”biologi” ja ”säädösteksti”.

Tässä yhteystavassa pitää huomioida erityisesti muodot kuten ”maa- ja metsätalousministeriö”, ettei pelkkä yhdyssanan määriteosa yhdisty toiseen sanaan. Näistä muodoista ei tehdä sanayhteyttä, muutoin esimerkiksi sanan ”metsätalousministeriö” 32:sta yhteydestä 14 olisi sanaan ”Maa” ja 15 olisi sanaan ”maa”. Muiden yhteyksien ollessa ”kasvituotanto” ja ”ympäristöministeriö” kaksi kertaa.

Käyttämällä mallia **”adjektiivi kuin substantiivi”** saadaan Wikipediasta yli 8000 osumaa ja mallilla **”yhtä adjektiivi kuin substantiivi”** saadaan reilut 600 osumaa. Näissä voi tosin sama sanapari esiintyä useamman kerran. Kun näissä huomioidaan vaatimus sanojen samasta taivutusmuodosta, putoavat luvut noin 6400:aan ja noin 470:een. Saatuihin yhteyksiin kuuluu sanojen ”kevyesti” ja ”tuulenhenkäys” yhteys, sekä neljä osumaa saanut yhteys ”teräsmies - voimakas”. Myös yhteys ”sulavasti - DOS” saa kaksi osumaa. Joissakin kuin vertauksissa verrataan toiseen samanlaiseen asiaan käyttäen adjektiivia, joka ei kuitenkaan kuvaa itse kyseessä olevaa asiaa hyvin suhteessa kaikkiin muihin samanlaisiin asioihin. Esimerkiksi kirjan ”Salaperäinen saari” tekstin osasta ”suunnilleen yhtä suuri kuin Malta” saadaan ”Maltan” ja ”suuren” yhteys, vaikka ”Malta” on hyvin pieni valtio tai saari. Tämä onkin suurin heikkous tässä mallissa. Joissain tapauksissa vertauksessa oleva kaksisanainen ilmaus tuottaa ongelmia. Esimerkiksi ilmaus ”yhtä vanha kuin purjehduksen historia” tuottaa yhteyden sanojen ”vanha” ja ”purjehdus” välille. Tämä toki korjaantuu monin paikoin kun sanojen vaaditaan olevan samassa muodossa.

Saman mallin hieman erilainen versio **”adjektiivi(Comp) kuin substantiivi”**, jossa adjektiivi on komparatiivi, tuottaa monia huonolta vaikuttavia yhteyksiä. Esimerkiksi ilmaus ”etenemisnopeus on suurempi kuin lämpöliike aineessa” on tuottanut ”suuri” ja ”lämpöliike” sanojen yhteyden. Myös yhteys sanojen ”taistelupanssarivaunu” ja ”kevyt” välillä vaikuttaa huonolta ja sen huomataan muodostuvan lauseesta ”Rynnäkköpanssarivaunun panssarointi on huomattavasti kevyempi kuin taistelupanssarivaunun”. Muita näin saatuja huonolta vaikuttavia yhteyksiä ovat sanan ”suuri” yhteydet muiden muassa sanoihin ”Yhdysvallat”, ”Englanti”, ”Alankomaat”, ”lehmä”, ”matkapuhelin” ja ”jättietana”. Ilman komparatiivia sana ”suuri” yhdistyy muiden muassa sanoihin ”Itävalta”, ”Malta”, ”Teksas”. Nämä ovat kaikki varmasti tietyissä tapauksissa suuria, mutta keskinäisissä vertailuissa näin ei kuitenkaan ole. Mallin komparatiivimuodolla saatuja paremmin onnistuneita yhteyksiä edustavat

sanojen ”jääkausi” ja ”viileä” yhteys sekä mahdollisesti myös tulkinnanvarainen yhteys ”vähittäiskauppa” ja ”kallis”. Sen sijaan yhteys sanojen ”kaularanka” ja ”yleinen” välillä osoittaa että sana ”yleinen” ei ehkä sovellu hyvin sanayhteyksiin.

Mallilla **”substantiivi on/oli/ovat/olivat adjektiivi”** saatiin kerättyä yli 160 tuhatta sanaparia. Näissä on monia hyvältä vaikuttavia yhteyksiä, mutta myös monia huonoja. Esimerkiksi ”kidutus” yhdistyy sanaan ”kielletty” kolme kertaa, ”pizzapohja” yhdistyy sanaan ”ohut” ja sana ”pikkuvaltio ” yhdistyy sanaan ”jatkuvasti”. Sanojen tunnistusta käsiteltäessä mainittu esimerkkisana ”e-hydroksidimetyyliallyylipyrofosfaatti” yhdistyy tällä mallilla sanaan ”orgaaninen”. Tällä mallilla selviää myös että ”vastaus” voi olla ”kielteinen” (5 osumaa), ”piilotettu” tai ”tyly” (molemmat 2 osumaa), myös sanat ”ehdoton”, ”nopea”, ”yksiselitteinen”, ”seuraava”, ”ystävällinen”, ” epäselvä”, ”kauhea”, ”lakoninen”, ”epätyydyttävä”, ”jyrkkä”, ”sisilialainen”, ”ilmeinen”, ”selvä”, ” lyhyt” ja ”helppo” saavat yhden osuman.

Mallilla **”substantiivi on/oli/ovat/olivat substantiivi”** yritettiin saada kerättyä kategoriatyyppisiä yhteyksiä, kuten esimerkiksi että ”kissa” on ”eläin” ja ”vasara” on ”työkalu”. Tämä ei kuitenkaan onnistunut kovinkaan hyvin. Saaduissa yhteyksissä esimerkiksi sana ”eläin” yhdistyy useisiin sanoihin muttei juurikaan eläimiin, kuitenkin sana ”ihminen” saa muutaman osuman sanaan ”eläin”. Sen sijaan sanoille ”työkalu” ja ”ase” saadaan vahva yhteys, kuten myös sanoille ”keskikenttä” ja ”pelipaikka”. Myös sanat ”petoeläin” ja ”jänis” saavat yhteyden, kuten saavat ”Einstein” ja ”professori” sekä ”Hitler” ja ”maalari”. Monet saaduista yhteyksistä yhdistivät erisnimen ja toisen sanan, useat näistä ovat ”kaupunki” ja ”kaupungin nimi” pareja.

Mallit **”substantiivi tai substantiivi”** ja **”adjektiivi tai adjektiivi”** tuottavat joitain hyviä ja joitain omituisia yhteyksiä. Esimerkkinä yhteydet ”fantastinen - maaginen”, ”kopiokone - pakokaasu”, ”urheiluauto - omakotitalo”, ”mopoauto - formula”, ”varkaus - vahingonteko / kavallus / murha / näpistys / ryöstö / petos”, ”pullapitko - astiankuivausteline” ja ”kookas - pieni (2kpl)/ muotoinen / keskikokoinen (9kpl)”. Tällä mallilla saatu hieman oudolta vaikuttava pakokaasun ja kopiokoneen välinen yhteys muodostuu tuoksuherkkyyteen liittyvästä lauseesta ”...kumin, pakokaasun tai kopiokoneen haju.”, jossa pitäisi osata tulkita yhteyden olevan pakokaasun hajun ja kopiokoneen hajun välillä.

Mallin **”yhtä adjektiivi ja adjektiivi kuin”** tulokset vaikuttavat erittäin hyviltä, vaikka niitä onkin hyvin vähän (44 kpl Wiki ja 11 kpl PG). Näitä löytyy sanamäärään suhteutettuna huomattavasti enemmän PG-teksteistä, eli lähes 13 kertaisesti Wikipediaan nähden. Ainoastaan yhteys ”ketterä - nopea” saa kaksi osumaa ja kaikki

muut tämän mallin yhteydet saavat vain yhden osuman. Muita yhteyksiä ovat muiden muassa ”massiivinen - raskas”, ”kylmäverinen - julma”, ”vehreä - puistomainen”, ”tiukka - ankara”, ”hyväntahtoinen - naiivi”, ”nopeasti - voimakkaasti”, ”huonokuntoinen - sairas”, ”monimutkainen - lento-pitoinen” ja ”leveä - korkea”.

Mallilla **”adjektiivi mutta adjektiivi”** toivottiin löydettävän vastakohtia. Näitä mallin mukaisia ilmaisuja on aineistossa yhteensä 675kpl, kun huomioidaan sanojen sama taivutusmuoto. Kun adverbien ja adjektiivien ei anneta olla toistensa pareina tässä muodossa putoaa ilmausten määrä 593:een. Mallilla löydettiin muiden muassa ilmaisut ”kapea mutta pitkä”, ”leveä mutta matala”, ”tyly mutta hyväsydäminen” ja ”hitaasti mutta varmasti”. Osa ilmaisuista sopisi hyvin vastakohtien muodostamiseen, mutta osasta taas saisi paremmin ominaisuuden naapuriominaisuuksia.

Vastakohtia kokeiltiin myös kerätä **”adjektiivi vaikka adjektiivi”** mallia käyttäen. Tällä toivottiin löydettävän ilmauksia kuten esimerkiksi ”iloinen vaikka sairas”. Tämä ei kuitenkaan tuottanut hyviä tuloksia ja kaikenkaikkiaan saatiin vain 35 osumaa, joista 30 täytti saman sanamuodon vaatimuksen. Vain muutamat näistä sanapareista vaikuttivat vastakohdilta, esimerkiksi ”tavallinen-satunnaisesti” ja ”hengissä-heikko” sanaparit. Yksi lupaavalta näyttävä, tosin ehkä hieman epäkorrekti, ilmaus oli perusmuotoisilla sanoilla ”pätevä vaikka liettualainen”. Tämä ei kuitenkaan täytä saman sanamuodon vaatimusta. Kun lisäksi huomioidaan tämän ilmauksen tuottava lause ”...vaalit olisivat päteviä vaikka liettualaiset eivät niihin jostain syystä osallistuisi.” niin selviää, ettei kyseessä ole vastakohtaa tarkoittava sanapari.

Myös muotoa **”melkein/lähes/jokseenkin yhtä adjektiivi kuin substantiivi”** kokeiltiin, mutta sillä saatiin vain muutama osuma koko aineistosta.

Vastakohtia yritettiin kerätä myös muodolla **”adjektiivi (superl) adjektiivi (superl)”**, jossa sanat ovat superlatiiveja. Esimerkiksi ”hitaimmasta nopeimpaan”. Tällaisia ei kuitenkaan löytynyt aineistosta ollenkaan.

3.3.2 Havaintoja Wikipedian ja Project Gutenberg -tekstien sanayhteyksistä

Koska vierekkäisten sanojen käyttäminen yhteyksien muodostamiseen ei välttämättä ole hyvä keino sanojen yhteyksien määrittämiseen ja muillakin yhteyksien muodostustavoilla saadaan yksittäisiä outoja yhteyksiä, käytetään yhteyksissä painotuksia/pisteytystä niin, että vierekkäisten sanojen välinen yhteys on pisteen arvoinen ja muut yhteyksien muodostustavat ovat kahden pisteen arvoisia. Myöhemmin (sa-

nayhteyksiä luettaessa ironian tuottamisvaiheessa) voidaan vain yhden (tai useammankin) osuman/pisteen saaneet yhteydet pudottaa pois. Näin jäljelle jäävät vahvimmat useimmin esiintyvät yhteydet.

Edellä kuvatuilla sanayhteyksien muodostusmenetelmillä saadaan tuotettua taulukon 1 mukaisesti yhteyksiä. Verrataan eri tapojen toimivuutta Wikipedian ja Project Gutenberg otannan kanssa. Wikipedia sisältää sanoja noin 50 miljoonaa ja Project Gutenberg tekstit noin miljoona sanaa. Vierekkäisten sanojen yhteyksiä ”adjektiivi adjektiivi”, ”adjektiivi substantiivi” ja ”substantiivi substantiivi” löytyy Wikipediasta noin kaksinkertainen määrä PG-teksteihin nähden suhteutettuna aineistojen kokonaissanamäärään. Myös ”substantiivi on/oli/ovat/olivat adjektiivi” ja ”substantiivi ja substantiivi” yhteyksiä on Wikipediassa huomattavasti enemmän kuin PG-teksteissä. ”adjektiivi ja adjektiivi” yhteyksiä on PG-teksteissä 1,5 kertaisesti wikipediaan nähden. Muissa yhteyksissä määrät ovat hyvin pieniä, mutta ”adjektiivi kuin substantiivi”, ”yhtä adjektiivi ja adjektiivi kuin” ja ”yhtä adjektiivi kuin substantiivi” malleja on tiheämmin PG-teksteissä kuin Wikipediassa.

Kaikkia edellä esiteltyjä yhteyksien muodostusmenetelmiä käytettäessä saatuja tuloksia tutkimalla havaittiin, että nimien käyttö yhteyksissä tuottaa lähinnä vain huonoja yhteyksiä. Nimet yhdistyvät toisiinsa ja muutoin melko satunnaisesti muihin sanoihin. Tunnetut merkkihenkilöt eivät yhdistyneet erityisen vahvasti mihinkään mielenkiintoisiin sanoihin ja saivat monia vähemmän oleellisia yhteyssanoja. Esimerkiksi Hitler on stereotyyppisesti kuollut ja liittyy läheisesti sanaan ”kansallissosialisti”. Nämä ovat siis eniten yhteyksiä saaneet sanat, monet muut yhteyksiä saaneet sanat vaikuttivat huonoilta yhteyksiltä, kuten ”noussut” ja ”suunnitelma”. Monet nimiin yhdistyvät adjektiivit ovat huonosti asiaa kuvaavia kuten esimerkiksi ”myöhemmin”, ”paljon” ja ”ennen”. Myös Newton, Einstein, Stalin ja Kekkonen saavat joitakin hyviä sanayhteyksiä, mutta enemmistö yhteyksistä vaikuttaa huonoilta. Nimi Matti yhdistyy huomattavan vahvasti nimiin Liisa ja Teppo, molempien saadessa yli 120 yhteyspistettä (eli ainakin yli 60 yhteyttä), kun seuraavaksi yleisin yhteys nimeen Mikko jää 22 pisteeseen. Matin ja Liisan yhteys selittynee aineistossa olevalla Juhani Ahon Rautatie teoksella, jonka päähenkilöitä Matti ja Liisa ovat.

Nimien poistaminen niin, että kaikki sanat muutetaan pienikirjaimisiksi, vähentää noin 10% yhteyksien ja erilaisten sanojen kokonaismäärää. Pienellä kirjoitettuja nimiä ei pääsääntöisesti ole tunnistettu sanoiksi, poikkeuksena luonnollisesti nimet kuten ”Satu”. Eniten vähenevät substantiivien keskinäiset yhteydet ja jotkin adjektiivien yhteydet kasvavat hieman, koska jotkin sanat käsitellään tällöin adjektiiveina

Yhteystapa	Wiki	PG	Sanojen määrä	N-grammit
”S on A”	213522	1452	51917	872071
”S on S”	102259	312	39911	354324
”A kuin S”	6431	358	4149	25841
”yhtä A kuin S”	474	35	533	4787
”yhtä A ja A kuin”	44	11	97	646
”A vaikka A”	28	2	53	443
”A tai A”	9625	67	3522	123760
”S tai S”	33387	184	20965	743853
”A A” vierekkäin	431182	8071	17241	2094225
”S(gen) S(nom)” vierekkäin	862408	4897	126681	10498309
”A S” vierekkäin	2281139	28382	198039	39198553
”A ja A”	59458	1746	9080	1487249
”S ja S”	408370	3509	94802	5080208
”A mutta A”	586	7	582	19336
Kaikki tavat	4306654	48721	293500	60503605

Taulukko 1: Yhteysmäärät eri tavoilla Wikipedialle ja Project Gutenberg (PG) teksteille eriteltynä, ehdolla että sanamuodot ovat samat yhteyksissä. A=adjektiivi, S=substantiivi. Sekä näissä yhteyksissä esiintyvien yksittäisten sanojen määrä Wikipedia ja Project Gutenberg aineistosta. Vertailun vuoksi yhteysmäärät myös n-grammeilla(nimet ja adverbit eivät ole mukana).

eikä substantiiveina.

Kaikkien yhteysmenetelmien yhteistuloksia tarkastelemalla havaitaan, että monet sanayhteydet saavat vain yhden yksittäisen osuman ja jotkin sanat yhdistyvät todella moniin sanoihin. Jotkin näistä paljon yhteyssanoja saavista sanoista eivät ole erityisen hyvin ominaisuuksia kuvaavia, kuten esimerkiksi sanat ”jo”, ”ensimmäinen”, ”oma”, ”muu” ja ”uusi”. Eniten yhteyksiä saavilla sanoilla voi olla useita kymmeniä tuhansia yhteyksiä ja ne voivat yhdistyä jopa kymmeneen tuhanteen eri sanaan. Kuitenkin löytyy myös joitakin mielenkiintoisia yhteysrakenteita, kuten esimerkiksi, että tyttöystävään liittyy vahvasti muiden muassa sanat entinen, silloinen ja pitkäaikainen. Avovaimoon laillinen ja sihteeri, vaimoon sanat kaunis sekä ystävä. Ex-vaimon naapurustoon kuuluvat sanat teini-ikäinen ja tytär ja aviovaimoon liittyy sanat äiti, lapsi ja diktaattori. Sen sijaan aviomies on vahvimmin yhteydessä sanoihin uusi tai entinen, poikaystävä on entinen, ex-poikaystävä on toimittaja ja robottipoikaystävä

on täydellinen.

Vealen sanayhteyksien laadukkuudesta ei ole tietoa, mutta hän antaa yhden esimerkin sanasta ”baby”, joka saa 163 stereotyyppistä ominaisuutta kuten ”baptized, adopted, toothless, soft, drooling, cute, hairless, pink, cuddly, fat, young, fresh, loved, demanding, chubby, happy, crying” [Vea12]. Vastaavasti sanalle ”vauva” saadaan 36 stereotyyppistä ominaisuutta, muiden muassa seuraavat ”pieni, kuollut, vastasyntynyt, uusi, iso, oikea, ikäinen, syntynyt, vaikeahoitoinen, suurisilmäinen, orpo, pullea, vihreä-hiuksinen ”. Sivuhuomautuksena mainittakoon, että ”demonivauva” on stereotyyppisesti ”punainen”.

3.3.3 Sanayhteydet n-grammeista

Sanayhteyksien muodostamiseen n-grammeista käytetään samoja malleja kuin edellä. Samoin vaaditaan myös sanojen sanamuodon olevan sama, ellei toisin mainita. Tulokset ovat hyvin samansuuntaisia kuin edellisen luvun Wikipedia ja Project Gutenberg aineistolla.

Vierekkäisistä sanoista muodostettavat yhteydet eivät toimi tässäkään erityisen hyvin. Lisäksi yhteyksiä saadaan erittäin paljon, vaikka nimiä ei hyväksytä mukaan yhteyksiin. Sanapari **”adjektiivi substantiivi”** tuottaa lähes 40 miljoonaa osumaa, **”adjektiivi adjektiivi”** yli 2 miljoonaa ja **”substantiivi substantiivi”** yli 10 miljoonaa osumaa, vaikka substantiivi parin sanojen vaaditaan olevan genetiivi ja nominatiivi. Erilaisia sanoja on pelkästään ”adjektiivi substantiivi” yhteyksissä yli 740 tuhatta. Saatuja yhteyksiä ovat esimerkiksi ”nautanahka - korkealaatuinen / rapea / purkukestävä / pehmeä / kova / italialainen”, ”piraattihassuttelu - itsestäänselvä”, ”peltikasa - laho / iso / ruosteinen”, ”korttivoi - luvallinen”, ”huumepilleri - lukuinen”, ”teiniuho - tavanomainen”, ”tietoturvaskanneri - näppärä”, ”sakaalimainen - eläin”, ”materiaalisammio - ehtymätön”, ”viljelmäkohtainen - sertifikaatti / arviointi”, ”haminakurittaja - kokenut”, ”koillistuuli - leuto / armoton / puuskittainen / kauhea / kova / rivakka / voimakas / heikko / kylmä / nasakka / hyinen / navakka / kuiva / kolea”, ”peruuttaaedellytys - kohta / oleskelulupa”, ”betonipintainen - rouhea / pyöreä / leveä”, ”kahvitaukoaika - lipunmyyjä”, ”mtstiedosto - muuntaa”, ”hiihtovalmennuslinja - kehittää”, ”kakkutoiminto - keitin”, ”ysimillinen - jäinen / entinen / pelkkä”, ”9001pohjainen - iso”.

Monet vierekkäisten sanojen sanayhteyksillä saadut sanaparit sopivat hyvin yhteen, erityisesti adjektiivi-substantiivi pareissa, mutta eivät kuitenkaan aivan toivotulla

stereotyyppistä ominaisuutta kuvaavalla tavalla. Lisäksi saatuja erilaisia yhteyksiä on hyvin suuri määrä, mutta suuri osa niistä on kuitenkin suhteellisen harvoin esiintyviä.

Käytettäessä **”ja”** sanaa yhdistämään kaksi substantiivia tai kaksi adjektiivia toisiinsa, saadaan n-grammeistakin monia hyviä yhteyksiä. Sanojen ”lääkäri” ja ”sairaanhoitaja” välille saadaan tässä erittäin vahva yhteys(yli 2000 pistettä, kun useat muut yhteydet saavat vain noin 10 pistettä). Huomattiin, että tälläkin aineistolla **”substantiivi ja substantiivi”** yhteyksissä ovat monikossa olevat yhteydet parempia kuin yksikössä olevat. Monikossa saadaan esimerkiksi yhteydet ”huippukännykkä - pokkari”, ”untuvatakki - untuvahousu / villapaita / pilkkihaalari / reppu / hame /”, ”satulahuopa - loimi / panta / pinteli /harja / riimu / satulavyö / suoja / juhlasatula / ohja”, ”dementoitua - sairastelu”, ”aamuvoimistelu - uinti”, ”käsidesipullo - sähköovi”, ”raparperipalo - sitruunaviipale”. Sanojen ollessa yksikössä saadaan muiden muassa yhteydet ”yhteisbiisi - taistelu”, ”laitteistotehokkuus - lämmöneristys”, ”nautanahka - pukinnahka”, ”kitararaiskaus - tehottomuus”, ”maaseutupäällikkö - tie”, ”aitalauta - sähkölanka ”, ”turvaistua - vauva”, ”koillistuuli - koleus / sade”, ”kylpyhuonelika - saippua”.

Mallilla **”adjektiivi ja adjektiivi”** saadaan seuraavanlaisia yhteyksiä: ”valkootsainen - korkeaääninen ”, ”kuolemavaarallinen - tavanomainen”, ”mukautumiskykyinen - tarkoituksenmukainen / joustava / moninainen / tilava”, ”korkeaseosteinen - ruostumaton”, ”astetarkka - nopea”, ”pienivastuksinen - kevyt”. Sana ”joustava” saa yli 8682 yhteysosumaa (luvusta ei voi päätellä paljoakaan tarkasti, mutta se kertoo kuitenkin, että sana saa hyvin paljon yhteyksiä). Joitakin näistä yhteyksistä ovat(luku kertoo yhteyden pisteet, jotka ovat keskenään verrannollisia suhteessa esiintymismääriin): ”kimmoinen 16, muuntua 12, puolueeton 12, tasokas 28, vastuuntuntoinen 12, saumaton 36, kimmoisa 208, monikäyttöinen 24, irtonainen 44, parempi 60, stressivapaa 12, hauska 24, kätevä 220, kestävä 1054, silkinpehmeä 20, rento 388, merkittävä 12, käytännönläheinen 66, ripeätoiminen 12, muodikas 6, oleellinen 12, kiireetön 24, terävä 8”

Toisin kuin HFST:llä tulkituissa sanoissa, n-grammiaineistossa "maa- ja metsätalousministeriö"muotoisista ilmauksista ei voida päätellä, että ensimmäinen osa on katkaistu yhdyssana. Tämä johtaa joihinkin huonoihin yhteyksiin. Esimerkiksi sana ”markkinointimies” yhdistyy sanaan ”nainen”, vaikka sen voisi olettaa tarkoittavan sanaa ”markkinointinainen”. Wikipedia ja Project Gutenberg aineistolla sanan ”metsätalousministeriö” yhteyksiä hallitsi sana ”maa”. Tässä ”maa” saa noin neljä kertaa

vahvemman yhteyden kuin useat muut yhteyssanat, mutta vahvin yhteys syntyy sanaan ”ympäristöministeriö” ja yhteys sanaan ”maaseutuvirasto” on myös huomattavasti vahvempi kuin sanaan ”maa”.

Käyttämällä mallia **”adjektiivi kuin substantiivi”** ehdolla, että adjektiivi ei ole komparatiivi muodossa saadaan esimerkiksi yhteydet: ”räiskintäpelaaja - enempi”, ”hannibal - vertahyytävä”, ”parisuhde - monenlainen”, ”sievä - karamelli / nenänpää / nukke / sika”, ”paistinpannu - kuuma”, ”rantakimma - kaunis”, ”puhdasrotuinen - risteytyskani”, ”dracula - kiehtova”. Monet näin saaduista yhteyksistä vaikuttavat hyviltä, osa on silti hyvin epämääräisiä.

Mallilla **”yhtä adjektiivi kuin substantiivi”** saadaan myös monia hyviltä vaikuttavia yhteyksiä. Sanan ”yhtä” mukanaolo mallissa ei näytä vaikuttavan merkittävästi yhteyksien laatuun verrattuna malliin ilman sanaa ”yhtä”. Myös tässä on ehtona että adjektiivi ei ole komparatiivi muodossa. Esimerkiksi seuraavat yhteydet ”kireä - kokoomus”, ”juna - nopea”, ”norsulauma - huomaamaton”, ”kansa - pienituloinen”, ”sopulilauma - hysteerinen”, ”kookospähkinä - älykäs”, ”nurmes - velkainen”.

Mallilla **”substantiivi on/oli/ovat/olivat adjektiivi”** saatiin joitain hyviä ja joitain huonoja yhteyksiä. Esimerkiksi seuraavat: ”hitausmassa - mystinen”, ”hansonlook - kova”, ”superraikas - sitruuna”, ”linja-auto - tärkeä /hyvä / suosittu / kätevä / 13metrinen / paras / ainoa / helppo / nykyaikainen”, ”kosketusherkkä - rintalihas / taulu / näppäinosa / ruutu / pinta / kosketusnäyttö”, ”maantieto - luotettava”, ”roikkumassa - paras”. Sanat ”tärkeä” ja ”suuri” saivat näin erittäin paljon yhteysosumia, 19464kpl ja 28650kpl.

Mallilla **”substantiivi on/oli/ovat/olivat substantiivi”** ei saada paljoa hyviä yhteyksiä, enemmänkin oudolta vaikuttavia. Esimerkiksi yhteydet ”bussikyyti - lapsi”, ”lepakkoryhmä - joukkue”, ”linja-auto - osa / pakko / linjaliikenne / kulkuväline / pysähdys / vaihtoehto”, ”museointi - katastrofi”, ”ilmaisnetti - puute”, ”roikkumassa - taskulamppu”.

Mallit **”substantiivi tai substantiivi”** ja **”adjektiivi tai adjektiivi”** tuottavat joitain hyviä, mutta myös monia huonoja yhteyksiä. Osa yhteyksistä on hieman vastakohtamaisia. Esimerkiksi adjektiiveille saadaan yhteydet ”lyhetä - pidetä”, ”pintaviallinen - pieni”, ”matelu - kiihdytys”, ”80vuotias - ckuppirintainen / vanha / 15vuotias”. Substantiiveille saadaan yhteydet ”helle - kylmä / kova / jäätävä”, ”kierrätyskelpoinen - polttokelvoton”, ”lempisarja - lempileffa / elokuva”, ”kymppitonni - tonni / satanen / kymppi”, ”aivopesu - ryhmäkuri / hallita / epäloogisuus / hypnoosi / provokaatio / manipulaatio / nuoleskelu / uskonnonopetus / pörssikeinottelu”, ”ke-

säkunnostus - poisto", "huoneilma-aukko - ulkoilmaventtiili", "johtosotku - kaapeli".

Mallin **"yhtä adjektiivi ja adjektiivi kuin"** tulokset vaikuttavat suhteellisen hyviltä. Saadaan esimerkiksi yhteydet "suojaisa - lämmin", "mutkainen - kapea", "avoin - ristiriitainen / vilpitön / vapaakulkuinen", "yksilöllinen - harvinainen / ainutlaatuinen", "omaperäinen - älykäs / fiksu / vetävä", "junttimainen - surullinen", "kiinnostava - tekninen / hikinen / kaunis".

Mallilla **"adjektiivi mutta adjektiivi"** saadut tulokset eivät ole erityisen hyviä. Yhteyksissä on jonkin verran vastakohtamaisuutta, muttei suoria vastakohtia. Mallilla löydettiin muiden muassa yhteydet: "asiallisesti - perusteellisesti / rennosti / rehellisesti / vapaasti / kuivasti / lyhyesti / napakasti / sitten / kepeästi / jämäkästi / osin", "oikeudenmukainen - julma / korkea / väkivaltainen / vaativa / ankara / tiukka / suuri", "kohtelias - vaisu / terävä / tyylikäs / pidättyväinen / viileä / etäinen / veijarimainen", "sateinen - lämmin / suloinen / leuto / lyhyt / mukava", "sievä - pieni / hauras / keltainen / huomaamaton".

Mallilla **"adjektiivi vaikka adjektiivi"** saadut yhteydet eivät ole erityisen hyviä tai huonoja. Yhteyksistä esimerkkeinä "hassu - hauska", "poutainen - pilvinen", "vaaraton - kivulias / kiusallinen" ja "asiallinen - ohutpeltinen".

Muodolla **"melkein/lähes/jokseenkin yhtä adjektiivi kuin substantiivi"** saatiin 148 yhteysosumaa ja 231 eri sanaa. Sana "jokseenkin" ei esiinny näissä ollenkaan. Monissa tapauksissa adjektiivina on "sti'-päätteinen adverbi. Esimerkkeinä seuraavat yhteydet "betoni - kova", "taksi - nopeasti", "animehahmo - kummallisesti ", "huussi - suuri", "ihmiskunta - vanha", "rakas - olut", "leggingsi - tiukka", "karmea - helvetti".

Muodolla **"adjektiivi(superl) adjektiivi(superl)"** saatiin 943 yhteysosumaa, mutta vain 208 eri sanaa. Esimerkkeinä seuraavat yhteydet "innokas - uusi", "kirjallinen - aukoton", "aktiivinen - pieni", "kuuma - kylmä / kuiva", "ovela - suuri", "varhainen - merkittävä / tunnettu", "lyhyt - ilmava / mahdollinen", "metallinen - korkea", "sekainen - läikikäs", "nopea - luotettava / nopea / rytmikäs / uusi / mahdollinen", "todellinen - syvä", "diagnostinen - tavallinen", "mukimitallinen - summittainen".

N-grammeista myös tutkittiin erikseen adverbejä sanayhteyksissä. Tätä ei tehty Wikipedia ja Project Gutenberg aineistoilla, koska niistä saatiin yhteensä vain noin 2000 adverbiyhteysosumaa, joista vajaa 300 on useamman kuin yhden osuman saaneita yhteyksiä. Adverbejä käsitellään yhteyksissä kuin adjektiiveja, joten niillä ei ole omia sanayhteysmalleja, mutta ne ovat adjektiiviyhteyksien erikoistapauksia.

Havaittiin että adverbiyhteydet eivät olleet useimmiten erityisen hyviä, eikä niitä tullut erityisen paljoa.

Mallilla ”yhtä adverbi kuin substantiivi” saadaan 2581 yhteysosumaa, joissa on yksittäisiä sanoja 1256. Nämä yhteydet ovat kuitenkin hyvää tasoa. Esimerkkinä yhteys ”yhtä järjestelmällisesti kuin gradu”.

Mallilla ”adverbi kuin substantiivi” saadaan 8347 yhteysosumaa, joissa on yksittäisiä sanoja 2659. Näin saadut yhteydet ovat melko huonoja. Esimerkkinä yhteys ”katu - huonosti”.

Mallilla ”adverbi mutta adverbi” saadaan 2731 yhteysosumaa, joissa on yksittäisiä sanoja 304. Monet näistä yhteyksistä vaikuttavat kelvollisilta yhteyksiltä, mutta näissä on monissa kuitenkin hivenen vastakohtamaisuutta, muttei kuitenkaan niin, että niitä voisi ottaa suoraan vastakohdiksi. Näitä yhteyksiä ei kannata ottaa mukaan sanayhteyksiin. Esimerkkinä yhteys ”poliittisesti - puhtaasti”.

Mallilla ”yhtä adverbi ja adverbi kuin” saadaan 89 yhteysosumaa, joissa on yksittäisiä sanoja 106. Nämä yhteydet ovat hyviä. Esimerkkinä yhteys ”rationaalisesti - loogisesti”.

Mallilla ”adverbi vaikka adverbi” saadaan 51 yhteysosumaa, joissa on yksittäisiä sanoja 34. Nämä yhteydet eivät ole erityisen hyviä. Esimerkkinä yhteys ”nätisti - kimakasti”.

Mallilla ”adverbi tai adverbi” saadaan 5963 yhteysosumaa, joissa on yksittäisiä sanoja 698. Nämä yhteydet ovat usein kelvollisia, mutta osa on myös vastakohtamaisia. Esimerkkinä yhteys ”analogisesti - digitaalisesti”.

Mallilla ”adverbi ja adverbi” saadaan 83399 yhteysosumaa, joissa on yksittäisiä sanoja 1663. Näissä yhteyksissä on joitain hyviä ja joitain huonoja. Esimerkkinä yhteys ”mielenkiinnottomasti - hitaasti”.

3.3.4 Havaintoja n-grammiyhteyksistä

Saatuja yhteyksiä tutkittaessa huomattiin, että nimet eivät toimi yhteyksissä erityisen hyvin n-grammeillakaan ja nimien käytöstä luovuttiin. N-grammeista saaduissa yhteyksissä voi esiintyä joitakin huonoja yhteyksiä senkin vuoksi, että aineistossa voi olla, ja onkin, jo valmiiksi ironisia ilmauksia, tietokirjamaisessa Wikipediassa tämä on epätodennäköisempää. N-grammeista saadaan hyvin samankaltaisia tuloksia kuin Wikipediasta ja Project Gutenberg teksteistä saadut yhteydet. N-grammiaineiston

suuresta määrästä johtuen nämä yhteydet ovat usein kuitenkin huomattavasti vahvempia.

Esimerkkinä olleelle sanalle ”vauva” saadaan n-grammeista 85 stereotyyppistä ominaisuutta. Nämä saadut yhteydet vaikuttavat paremmilta kuin Wikipediasta ja Project Gutenberg teksteistä saadut yhteydet. Monet näistä n-grammiyhteyksistä ovat myös vahvempia. Esimerkiksi vahvimmat yhteyssanat ovat valmis, pieni, ihana, itkuinen, terve, helppo, nälkäinen, hyvä ja sairas. Nämä kaikki ovat huomattavasti vahvempia yhteyksiä kuin Wikipediasta ja Project Gutenberg teksteistä saatu vahvin yhteys sanalle ”vauva”.

N-grammiyhteyksissä voisi olla syytä huomioida yhteyden vahvuuden lisäksi se kuinka monesta erimallisesta rivistä yhteys tulee. Yhdestä rivistä, eli yhden tyyppisestä ilmauksesta, saatu vahva yhteys ei ehkä ole yhtä hyvä kuin useammasta rivistä saatu hieman heikompi yhteys. Tätä ei kuitenkaan tarkasteltu niin paljoa, että olisi päädytty selkeisiin johtopäätöksiin.

3.3.5 Valitut yhteysmenetelmät

Valitut menetelmät ja niillä saadut yhteysmäärät löytyvät taulukosta 2. Nimet jätettiin pois huonojen saatujen yhteyksien vuoksi. Jotkin erittäin paljon yhteyksiä saavat, mutta huonosti ominaisuutta kuvaavat sanat hylättiin kokonaan. Taulukon luvuista huomataan, että n-grammit dominoivat yhteyksiä eikä Wikipediasta ja Project Gutenberg aineistosta saaduilla yhteyksillä ole paljoakaan merkitystä.

Vierekkäisiä yhteyksiä käytettiin pääasiassa sana-/yhteysmäärien kasvattamiseen. Mutta, koska n-grammiaineistolla saadaan suuria määriä yhteyksiä muutenkin, niin niitä ei tarvita. Myöskään vierekkäisillä sanoilla käytetty malli ” substantiivi(nom) substantiivi (gen)” ei tuota tähän käyttöön parhaiten sopivia yhteyksiä. Sillä saadaan asia joka on toisen osa, kun yhteyteen haluttaisiin pikemminkin vastaavanlaiset asiat. Myös muut mallit, jotka eivät tuottaneet laadukkaita yhteyksiä jätettiin pois lopullisista valinnoista. Jäljelle jäivät seuraavat parhaimmilta vaikuttavat yhteyksien muodostus tavat.

Mallin ”substantiivi ja substantiivi” yhteyksistä otetaan huomioon vain monikossa olevat.

Muotoa ”(yhtä) adjektiivi kuin substantiivi” olevien yhteyksien huomattiin olevan parempia silloin, kun adjektiivi ei ole komparatiivi muodossa ja vain sellaiset yhteydet otetaan huomioon.

Malli ”substantiivi on adjektiivi” otettiin mukaan tuottamaan runsaasti substantiivien ja adjektiivien välisiä yhteyksiä, joita muutoin ei saada erityisen paljoa.

Mallilla ”adjektiivi ja adjektiivi” tuotetaan käytännössä kaikki adjektiivien keskinäiset yhteydet. Mallin ”yhtä adjektiivi ja adjektiivi kuin” tuottamat yhteydet ovat käytännössä aivan samoja, mutta nämä ovat pääsääntöisesti niin hyviä yhteyksiä, että ne otettiin mukaan vahvistamaan parhaita adjektiivien välisiä yhteyksiä.

Yhteystapa	Wiki	PG	N-gram
”S on A”	112288	735	873303
”A kuin S”	942	92	44034
”yhtä A kuin S”	241	19	4787
”yhtä A ja A kuin”	44	11	646
”A ja A”	52168	1457	1497079
”S ja S” monikko	77313	1007	1889855
Kaikki tavat	242996	3321	4318051

Taulukko 2: Lopulliset yhteysmäärät eri tavoilla Wikipedialle, Project Gutenberg teksteille ja n-grammeille eriteltynä. A=adjektiivi, S=substantiivi. Erot taulukkoon 1 selittyvät nimien pois jättämisellä ja joidenkin paljon yhteyksiä keräävien epämääräisten sanojen hylkäämisellä.

Kuten taulukosta nähdään, suurin osa yhteyksistä tulee n-grammeista. Wikipediasta saadaan yksittäisiä sanoja 54836 kpl, Project Gutenberg lisää yksittäisten sanojen määrää 482:lla (näitä sanoja ei siis ollut Wikipediayhteyksissä), n-grammeista saadaan suurin osa sanoista ja lisäystä aiempiin on 163630 uutta sanaa. Yhteensä erilaisia sanoja on siis 218948 kappaletta.

Tekstistä poimittujen sanayhteyksien lisäksi uusien ilmausten muodostusvaiheessa hyödynnetään erillistä listaa sanojen vastakohdista. Vastakohtalistalla olevia sanoja löytyy aineistosta(Wikipedia ja Project Gutenberg kaikilla yhteysmalleilla) tunnistetuista sanoista noin 950. N-gram-aineiston mukaanotto nostaa sanayhteyksistä löytyvien vastakohtien määrän noin tuhanteen. Taulukon 2 mukaisia yhteyksiä käyttäen saadaan sanayhteyksiin 1003 sanaa, joille löytyy vastakohta käytetystä vastakohtalistasta.

4 Tulokset

Lopulliset tulokset, joiden olisi tarkoitus olla ironisia, saadaan aikaan käyttämällä strategioita kuten Veale [Vea12]. Erona Vealen toteutukseen on, että tässä ei käytetä kategoriaoperaattoria, koska kategorioita ei saada poimittua aineistosta. Veale saa kategoriat poimittua suoraan WordNetistä. Wikipediasta näitä voisi saada osittain, sillä esimerkiksi eläinten ja kasvien luokittelut ovat usein artikkelin sivussa, mutta Wikipedian siivoamiseen käytetty skripti jättää pois näitä tietoja. Käytetyt yhteyksienmuodostusmenetelmätkään eivät juurikaan onnistu löytämään tekstistä selkeitä luokitteluja. Operaattoreina ovat siis stereotyyppioperaattori ”@”, naapurioperaattori ”?” ja vastakohtaoperaattori ”-”. Jos jonkin tietyn sanan kanssa esimerkiksi halutaan sanan naapurin vastakohta ja sanalle tyypillinen ominaisuus, tällöin strategiana on ”-? @”. Edellä oleva strategia tuottaa tuloksena kaksi sanaa, koska operaattoreiden välissä on välilyönti. Lisäksi käytössä on yhdyssanaoperaattori ”#”, jolla kaksi strategioiden tuottamaa sanaa voidaan yhdistää yhdyssanaksi. Tämä ei kuitenkaan takaa että muodostettava yhdyssana on oikea sana.

Alustavat testit vierekkäisten sanojen yhteyksillä: Pelkästään sanayhteyksillä, jotka on tehty sanojen vierekkäisen esiintymisen pohjalta, saadaan lupaavan oloisia järkeviltäkin vaikuttavia tuloksia. Tosin seassa on erittäin paljon myös outoja vaihtoehtoja. Näin kuitenkin olisi mahdollista saada monia Vealen tuloksia vastaavia tuloksia.

Esimerkiksi strategia ”@-P @-P” voi tuottaa sanasta ”pehmeä” tuloksen ”tiili seinämä/seinälaatta”. Veale sai muiden muassa tulokseksi ”brick/steel wall”.

Sanasta ”kova” saadaan strategialla ”?-P @-P” tulokseksi ”keitetty patja”, kun Veale sai vastaavasti tuotettua ”soggy pillow”.

Käytettäessä sanoja ”savage/villi” ja strategiaa ”?-P @?-P” Veale saa muodostettua ”meek lamb” kun tässä saadaan ”enkelimäinen lammas”.

Veale käyttää myös mallia ”as P as (?P @P)” tuottamaan mm. ”as cold as wet fish” sanasta ”cold”, vastaavasti ”kylmä” voi tuottaa vertauksen ”kylmä kuin hyytävä kala”.

Veale käyttää myös strategiaa ”@P @P”, jolla sanasta ”cold” saadaan ”robot fish” ja ”snow storm”. Suomeksi ”lumimyrsky” saadaan muodostettua jo yhdellä ”@P”:llä sanasta ”kylmä”, mutta pelkkä ”myrsky” ei yhdisty sanaan ”kylmä”. Näissä kaikissa vaihtoehdoissa on tosin hyvin paljon muitakin huonommin sopivia vaihtoehtoja.

Tässä onkin paljon merkitystä sillä mikä sana valitaan lopulliseen tuotokseen monien

mahdollisten joukosta ja millaisia sanayhteyksiä on saatu luotua.

Wikipedian ja Project Gutenberg -aineiston tulokset ja sanojen karsinta: Strategiat tuottavat usein monen sanan joukon, josta lopullinen tulos, eli yksittäinen sana, valitaan satunnaisesti. Näin valitsemalla saadaan tuloksiin vaihtelua, vaikka jotkut vaihtoehdot voivatkin olla hyvin outoja. Sillä, jos valinta tehtäisiin huomioimalla sanojen esiintymismääriä yhteyksissä suhteessa tiettyyn sanaan, voi joku sana saada huomattavasti suuremman lukuarvon kuin muut vaihtoehtoiset sanat ja näin se tulisi miltei aina tälle sanalle yhteydeksi. Erityisesti näin, jos strategia koostuu useammasta operaattorista ja ensimmäisellä operaattorilla jokin sana esiintyy huomattavasti useammin sanayhteyksissä kuin muut ja tästä sanasta seuraavalla operaattorilla saatavista sanoista on jokin huomattavasti useammin esiintyvä kuin muut. Näin saataisiin monta harvoin esiintyvää tulosta näistä harvoin esiintyvistä sanoista ja yksi usein esiintyvä liittyen usein esiintyvään sanaan. Esimerkiksi sana ”nopea” yhdistyy yli 500 adjektiiviin ja yli tuhanteen substantiiviin. Kuitenkin adjektiiveista vain alle puolet saa useamman kuin yhden yhteyden ja 50 kpl saa yli 10 yhteyspistettä. Kolmentoista vahvimman yhteyden saadessa yli 50 pistettä. Substantiiveistakin vain noin 200 sanaa saa useamman kuin yhden yhteysosuman ja 39 sanaa saa yli 10 yhteyspistettä. Sekä adjektiiveissa että substantiiveissa vahvin yhteys saa noin 160 yhteyspistettä. Eli vain pieni osa yhteyksistä saa huomattavan suuren osan kaikista yhteyspisteistä.

Käyttämällä kaikkia luvussa 3.3 esitettyjä sanayhteyksien muodostusmenetelmiä, saadaan sanojen yhteydet hieman paremmiksi. Kun myös pisteytetään yhteydet niiden laadukkuuden mukaan (esimerkiksi niin, että vierekkäinen yhteys on 1 pisteen arvoinen ja muut yhteydet ovat 2 pisteen yhteyksiä) ja jätetään vähän(alle 3) pisteitä saaneet sanayhteydet pois, jää monet huonoimmat sanayhteydet pois. Näin myös karsitaan vain yhden osuman saaneet yhteydet.

Esimerkkikäyttötapaus strategialle voisi olla tilanne, jossa halutaan ironinen muoto ilmauksesta ”nopea kuin gepardi” korvaamalla gepardi jollakin ei nopealla, eli hitaalla. Tällöin strategiaksi sopisi ”@-nopea”, jolla siis tuotetaan nopean vastakohdalle ominainen sana. Tässä voitaisiin toivoa saatavan tulokseksi esimerkiksi kilpikonna. Strategioissa on huomioitava, että pienetkin muutokset voivat vaikuttaa paljon. Tässä tapauksessa ”@-nopea” antaa saman tuloksen kuin strategia ”@hidas”, mutta toisinpäin eli strategioilla ”@-hidas” ja ”@nopea” tulokset ovat erilaiset. Tämä johtuu siitä että sanalla ”nopea” on vain yksi vastakohta, mutta sanalla ”hidas” on useampia vastakohtia. Strategialle ”@-nopea” saadaan seuraavia tuloksia ”nopeus”,

”räjähde”, ”kehitys”, ”vauhti”, ”balladi”, ”blues” ja ”kappale”. Samalle strategialle ”@- ” ja sanalla ”hidas” saadaan joitain hieman paremmilta vaikuttavia tuloksia ”sukellusveneentorjuntafregatti”, ”yllätyshyökkäys”, ”gepardi”, ”lentokone”, ”välimuisti”, ”internetyhteys” ja ”vinttikoira”.

Strategioita on helppo muokata pienillä muutoksilla. Esimerkiksi ”terävä” sanalle strategiat ”-?” ja ”?-” tuottavat täysin erilaiset tulokset, koska operaattorit ovat eri järjestyksessä. Näin erilaisia strategiavaihtoehtoja saadaan hyvin paljon pienillä muutoksilla. Strategia ”?-”, eli vastakohtien naapurisanat, tuottaa sanalla ”terävä” tulokseksi sanat ”tylsä”, ”liian”, ”kasvoton” ja erittäisn”. Sen sijaan strategia ”-?”, eli naapurien vastakohdat, tuottaa 50 eri sanaa, muiden muassa ”lyhyt”, ”laimea”, ”pehmeä”, ”kulunut” ja ”suunnaton”.

Testausmielessä kokeiltiin saako sanaa ”muoviveitsi”, tai jotain sitä läheisesti muistuttavaa, tuotettua sanasta ”terävä” strategialla ”@@-#@”. Tuloksiksi saadaan esimerkiksi ”leveä#kynä”, ”suuri#sarvi”, ”tylsä#sorkka”, ”hengellinen#mutka”, ”ainoastaan#neula” ja ”pitkä#peli”. Koska nämä tuloksina saadut sanat eivät muodosta oikeaoppisia yhdyssanoja, voitaisiin strategian yhdyssanaoperaattori korvata välilyönnillä ja tulokset muuttuisivat sanapareiksi.

Nyt useampia yhteysmalleja käytettäessä ja heikommat yhteydet karsittaessa saadaan aiemmin esitettyihin Vealen esimerkkeihin hieman erilaisia tuloksia.

Strategia ”@- @-” sanalla ”pehmeä” tuottaa joukon epämääräisiä tuloksia kuten ”alustava ääni”, ”paine verotus”, ”puuaines valuutta” ja ”vauhti tuuli”.

Strategia ”?- @-” sanalla ”kova” voi tuottaa esimerkiksi tulokset ”ohut turkki”, ”tiheä juusto”, ”joustava pohja” ja ”hauras patja”. Tässä monet mahdolliset vaihtoehdot vaikuttavat verrattain hyviltä.

Strategia ”?- @?-” sanalla ”villi” tuottaa tulokset ”lempeä ihminen”, ”ystävällinen turisti” ja ”äidillinen kirjallisuustiede”. Strategian jälkimmäinen osa ”@?-” tuottaa noin 3000 eri sanaa.

Strategia ”(yhtä kylmä kuin) ? @” sanalla ”kylmä” tuottaa tulokset ”kolkko pohjola”, ”talvisin napa-alue”, ”armoton erikoisala” ja vahvimmin yhdistyvin sanoin ”kuuma sota”.

N-grammiyhteydet mukaan: N-grammeista saadaan paljon vahvoja yhteyksiä, joten heikoimpia yhteyksiä voidaan karsia paljon enemmän kuin pelkkää Wikipedian aineistoa käytettäessä. Hyvä raja yhteyspisteille voi olla esimerkiksi 12 pistettä, joka vaikuttaa olevan satunnaisten ja useimmin esiintyvien yhteyksien raja-

na n-grammiyhteyksissä. Tällöin Wikipediasta ja Project Gutenberg -aineistoista saadut yhteydet karsiintuvat huomattavasti. Esimerkiksi sanan ”nopea” yhteyksistä vain 45 adjektiiviyhteyttä ja 35 substantiiviyhteyttä ylittävät 12 pisteen rajan. N-grammiyhteyksistä saadaan sanalle ”nopea” vastaavasti 814 adjektiivia 935:stä ja 541 substantiivia 1051:stä, joiden saamat yhteyspisteet ovat 12 tai enemmän. Sanan ”helppo” saadessa 79168 pistettä yhteyksistä sanaan ”nopea”. Substantiiveista vahvin yhteys, 320 pistettä, on sanalla ”toimitus”. Eniten yhteyksiä saavia sanoja tarkastelemalla huomataan, että karsimatta on jäänyt sana ”muu”, joka yhdistyy yli 7000 sanaan, näistä yli 5000 on yli 12 pisteen yhteyksiä.

Lopullisia, taulukon 2 mukaisia, parhaimmilta vaikuttavia sanayhteyksien muodostusmalleja käytettäessä saadaan sanayhteyksistä 218948 yksittäistä sanaa. Jos näistä karsitaan vain yhden osuman saavat sanat pois, jää jäljelle 182927 sanaa.

Nyt Vealen tuloksiin vertailtavat strategiat tuottavat seuraavanlaisia tuloksia.

Strategia ”@- @-” sanalla ”pehmeä” tuottaa outoja tuloksia kuten ”massaliike armeija”, ”taikina talvi” ja ”kilohinta lattia”. Suosituimpia, eli eniten pisteitä saavia, sanoja ovat ”yritys” (1338 pistettä), ”hinta” (1148 pistettä), ”vauhti” (718 pistettä), ”paine” (698 pistettä), ”taso” (690 pistettä)”. Useimmat tulokset näyttävät tähän sopimattomilta.

Strategia ”?- @-” sanalla ”kova” tuottaa tulokset ”värikäs iho”, ”luminen ääriviiva”, ”märkä lompakko”, ”empaattinen joustovinyylimatto ” ja ”kostea tyyny”. Suosituimpia sanoja ovat ”?-”-strategialle ”lämmin” (3636), joustava” (3300), sileä” (2266), kova” (2084) ja ”@-”-strategialle ”iho” (386), maku” (340), askel” (306), pinta” (194). Monet tulokset ovat huonoja, mutta myös useat mahdolliset tulokset vaikuttavat kuitenkin hyviltä.

Strategia ”?- @?-” sanalla ”villi” tuottaa tulokset ”ihmisystävällinen elokuva”, ”söpö lammas”, ”maltillinen työtilanne” ja ”näppärä päätös”. Strategia ”?-” tuottaa 247 sanaa(jotka ovat saaneet ainakin 12 yhteyspistettä), mutta strategia ”@?-” tuottaa 18082 sanaa. Suosituimpia sanoja ovat strategialle ”?-” ”rauhallinen” (1790), ”mukava” (660), ”ihana” (434), ”suloinen” (398) ja ”@?-”-strategialle ”lapsi (155340), ”tamma” (15737), ”ihminen” (14072), ”ori” (13676). Tässä monet saadut adjektiivit vaikuttavat hyviltä, mutta substantiivit selkeästi huonommilta.

Strategia ”(yhtä kylmä kuin) ? @” sanalla ”kylmä” tuottaa tulokset ”tunteeton tammikuu”, ”ankea jääkaappi”, ”kevyt kanahampurilainen”, ”huoneenlämpöinen morsian” ja ”syvä uuni”. Suosituimpia sanoja ovat strategialle ”?” ”lämmin” (8196), ”pimeä”

(5310), ”kuuma” (4610), ”luminen” (2806) ja strategialle ”@” ”maailma” (748), ”ilma” (374), ”talvi” (370), ”yö” (240). Sanan ”kylmä” naapuri sanat vaikuttavat melko hyviltä, mutta stereotyyppisesti kylmät sanat vaikuttavat oudoilta.

Myös tällä sanayhteysaineistolla kokeiltiin saako sanan ”muoviveitsi” tyylistä sanaa tuotettua sanasta ”terävä” strategialla ”@@-#@”. Tuloksiksi saadaan esimerkiksi ”voimaton#kirves”, ”harras#rauta”, ”kirjallinen#partaveitsi”, ”keskiaikainen#paino”, ”kaksiosainen#säilä”, ”ylittämätön#ilme” ja ”samansuuruinen#säilä”.

Havaintoja strategioilla tuotetuista sanoista: Näyttää siltä, että stereotyyppistä ominaisuutta (eli strategia ”@”) kuvaavat yhteydet eivät ole aina erityisen hyvin onnistuneita. Poikkeuksiakin on ja vaikuttaisi siltä, että vähemmän yhteyksiä saaville sanoille nämä yhteydet ovat usein parempia. Tästä johtuen kokeiltiin poistaa "substantiivi on adjektiivi”-mallilla, joka tuottaa suurimman osan adjektiivien ja substantiivien välisistä yhteyksistä, saadut sanayhteydet. Tämä auttoi joissain tapauksissa, mutta joissain tapauksissa yhteydet huonontuivat. Koska näin poistettiin suurin osa stereotyppiyhteyksistä, niin myös yhteyksien vahvuudet laskivat merkittävästi. Esimerkiksi sanalle ”pehmeä” strategia ”@-” tuottaa nyt vahvimpina sanat ”kivi”, ”kallio”, ”halu” ja ”ääni”. Sanalle ”kova” strategialla ”@-” saadaan vahvimpina ”pumpuli” ja ”poutapilvi” muiden (8 kpl), erittäin pehmeiden, yhteyssanojen saadessa 12 pistettä. Sanan ”villi” ja strategian ”@?-” tuloksiin tämä muutos ei näytä vaikuttavan paljoa, edelleen saadaan hyvin paljon oudolta vaikuttavia sanoja. Sanalle ”kylmä” strategialla ”@” saadaan nyt vahvimpina sanoina sanat ”jääkaappi”, ”vuosi” ja ”kivi”. Sanalle ”terävä” saadaan nyt strategialla ”@@-#@” satunnaiset tulokset ”kollegiaalinen#terä”, ”legendaarinen#partaveitsi” ja ”tylsä#miekka”.

Edellä esitetyt tulokset näyttivät parantuneen tällä "substantiivi on adjektiivi”-mallin yhteyksien poistamisella, mutta tämän muutoksen jälkeen esimerkiksi ”pöllö” ei ole enää ”viisas” vaan ainoastaan ”sokea” ja ”salaperäinen”. Monille vähän yhteyksiä saaneille sanoille tästä muutoksesta on selkeästi haittaa.

Aineisto (Wikipedia, Project Gutenberg tai n-grammit) ei näytä vaikuttavan lopulta merkittävästi tuloksiin, vaan enemmänkin vaikuttaa se millä malleilla yhteyksiä kerätään. Tuloksissa on samanlaiset sanat vahvimpien joukossa sekä Wikipedia että n-grammi aineistolla. N-grammeissa vahvimmat sanat ovat moninkertaisesti vahvempia ja yhteyssanoja on enemmän, mutta muutoin erot eivät ole lopulta kovinkaan suuria.

Yleisellä tasolla tuloksista huomattiin, että eläimet yhdistyvät hyvin toisiinsa, samoin kuin kulkuneuvot yhdistyvät hyvin keskenään. Hieman yllättävänä havaintona

huomattiin myös, että poliittisesti epäkorrektit sanat saivat yhteyslaadultaan melko hyviä, joskin epäkorrekteja, yhteyksiä. Tällaiset yhteydet näyttävät tulevan pelkästään n-grammeista.

Yksinkertaiset strategiat näyttävät olevan parhaita, useampia operaattoreita käytettäessä epämääräiseltä vaikuttavien sanojen määrä kasvaa usein hyvin paljon.

Yhteyssanoja ei kuitenkaan voi karsia vain isommalla hyväksyttävien yhteyksien kynnysarvolla, koska monet hyvät yhteydet eivät usein saa erityisen paljon pisteitä. Lisäksi nämä yhteyteen hyvin sopivat sanat eivät ole aina eniten pisteitä saavia. Monet huonommat yhteyssanat voivat saada suuria pistemääriä. Toisilla sanoilla on huomattavasti vähemmän yhteysosumia kuin toisilla ja yhteyksien vahvuuksissa on paljon sanakohtaisia eroja.

5 Yhteenveto

Toteutettiin prototyyppijärjestelmä, joka ”toimii kuin bussi”, mutta varsinainen ironian tuottaminen osoittautui hankalaksi.

Sanojen tunnistus toimii melko hyvin, ainakin tähän tarkoitukseen riittävästi. Käytetty menetelmä mahdollistaa melko helposti minkä tahansa normaalin tekstin käytön sanayhteyksien muodostusaineistona. Parannusta sanojen tunnistukseen voisi saada esimerkiksi ympäröivien sanojen tai koko lauseen huomioimisella.

Sanayhteyksissä käytetyt mallit ovat hyvin yksinkertaisia. Osa toimii hyvin, mutta parannettavaa on kuitenkin paljon. Mahdollisesti ympäristön sanojen huomioiminen voisi poistaa joitain ongelmia. Saaduissa sanayhteyksissä on monia hyviä yhteyksiä, mutta huonoimpien yhteyksien määrää pitäisi saada vähäisemmäksi. Malli ”substantiivi ja substantiivi” sanojen ollessa monikossa näyttäisi olevan selkeästi parhaiten toimiva sanayhteyksien muodostuksessa käytetyistä malleista.

Uuden ironisen ilmaisun tuottaminen osoittautui vaikeaksi. Strategiat toimivat hyvin ja ongelmat olivatkin enemmän sanayhteyksissä. Monet tuloksista ovat lupaavia ja voivat ainakin antaa käyttäjälle inspiraatiota luoviin ilmauksiin, muihinkin kuin vain ironisiin. Tuloksien laadussa on kuitenkin hyvin paljon sanakohtaisia eroja.

Seuraava askel tulosten parantamiseksi voisi olla lauseiden tarkempi analysointi ja käsittely kokonaisuutena sekä sanojen tunnistuksessa että sanayhteyksien luomisessa.

Loppuun mainittakoon ”gradun” olevan stereotyyppisesti vahvimmin ”valmis”.

Lähteet

Vea12 Veale, T. Detecting and Generating Ironic Comparisons: An Application of Creative Information Retrieval. *AAAI Fall Symposium Series 2012, Artificial Intelligence of Humor*, Arlington, Virginia, USA, sivut 101-108.

Vea11 Veale, T. Creative Language Retrieval: A Robust Hybrid of Information Retrieval and Linguistic Creativity. *Proc. of the ACL 2011, the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, sivut 278-287.

Vea04 Veale, T. The Challenge of Creative Information Retrieval. *The proceedings of CICLing Conference on Intelligent Text Processing and Computational Linguistics 2004*, Soul, Korea, sivut 457-467.

VeH10 Veale, T. ja Hao, Y. Detecting Ironic Intent in Creative Comparisons. *Proc. of ECAI'2010, the 19th European Conference on Artificial Intelligence*, Lissabon, Portugali, sivut 765-770.

VPM12 Voutilainen, A., Purtonen, T. ja Muhonen, K. (2012). FinnTreeBank2 manual. *University of Helsinki, Department of Modern Languages.* http://www.ling.helsinki.fi/kieliteknologia/tutkimus/treebank/sources/FinnTreeBankManual.pdf [20.9.2014]