hyväksymispäivä arvosana

arvostelija

Verkkotekstien kielentunnistus

Miina Kilpikivi

Helsinki 5.11.2010

Seminaariraportti

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Tietojenkäsittelytieteen laitos

Tekijä — Författare — Author
Miina Kilpikivi

Työn nimi — Arbetets titel — Title
Verkkotekstien kielentunnistus

Oppiaine — Läroämne — Subject

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Seminaariraportti	5.11.2010	11 sivua

Tiivistelmä — Referat — Abstract

Automaattinen kielentunnistus selvittää millä kielellä teksti on kirjoitettu. Usein automaatti toimii moitteettomasti ja tunnistaa kielet luotettavasti. Verkkotekstit ovat kuitenkin luoneet kielentunnistukselle haasteen, jota ei ole vielä monelta osin ratkaistu. Seminaariraportissa käydään läpi yleisesti käytettyjä kielentunnistusmenetelmiä, vaikeita tapauksia, joissa nämä menetelmät eivät toimi tarpeeksi hyvin ja näihin vaikeisiin tilanteisiin ehdotettuja menetelmiä. Lisäksi raportissa pohditaan verkkoteksteistä kerätyn korpuksen edustavuutta .

Avainsanat — Nyckelord — Keywords

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Sisältö

1 Johdanto

World Wide Web tarjoaa laajat mahdollisuudet kielentutkimukseen. Verkossa tekstit ovat helposti saatavilla, ilmaisia, tuoreita ja laaja-alaisia. Verkossa on valtavat määrät tekstejä eri kielillä, ja määrä kasvaa koko ajan. Verkkoteksteistä voidaan koota korpus, josta tehdään kielitieteellisiä analyysejä [KG03].

Ensimmäinen askel on selvittää mitä kieltä teksti edustaa. Korpusta kerättäessä tarvitaan automaattista kielentunnistusta, jotta tekstit edustaisivat nimenomaan sitä kieltä mitä halutaan kuvata ja jotta jokaisen tekstin kohdalla ei tarvittaisi ihmisen työtä kielentunnistamiseen. Kielentunnistus karkealla tasolla toimii hyvin, mutta monikieliset tekstit, hyvin lyhyet tekstit sekä toisiaan lähellä olevat kielet ja kielivariantit muodostavat haasteen. Tällöin on ongelmana, että koottu aineisto kuvaa vain automaatin tunnistamaa osaa kielestä ja näin aineisto sekä siitä tehdyt johtopäätökset vääristyvät. Ongelmana on myös tuntemattoman kielen luokittelu. Yleisesti käytössä olevat menetelmät tunnistavat tuntemattomankin kielen joksikin ehdokeistaan.

Yksinkertaisimmillaan automaattinen kielentunnistus voi hyödyntää sivun URL:ia verkkotekstin kielen selvittämiseen [BHW08]. Yleisempänä menetelmänä kielentunnistuksessa käytetään n-grammeja, jotka ovat erittäin tehokkaita ja virhesietoisia [MS05]. N-grammien lisäksi kielentunnistusta voidaan parantaa yleisten sanojen tai pienten sanojen listan käytöllä [SCH+94]. Monikielisten tekstien osalta kielentunnistus voidaan toteuttaa koko tekstin sijaan myös pienemmissä osissa, esimerkiksi kappaletasolla, jolloin saadaan eroteltua kielet toisistaan [Pra99]. Toisiaan lähellä olevien kielten tunnistamiseen myös niin sanottu kiellettyjen sanojen säännön (*rule of forbidden words*) käyttö on tuottanut luotettavia tuloksia [LMB07].

Korpuksen käytössä ja johtopäätöksien teossa on otettava huomioon korpuksen edustavuus [Bib93]. Edustaako korpus koko kielenkäyttöä, ja voidaanko sen pohjalta tehdä yleistyksiä kieleen liittyen vai edustaako se vain sitä joukkoa mistä se on koottu? Voiko ja onko sen tarkoituskaan kuvata muuta kuin itseään [KG03]?

Automaattinen kielentunnistus toimii sataprosenttisesti vain yksinkertaisissa tapauksissa. Vaikeassa kontekstissa tarvitaan kehittyneempiä menetelmiä, jotta saataisiin koottua luotettavaa aineistoa. Automaattista kielentunnistusta käytetään verkossa myös muihin tarkoituksiin, kuin korpuksen keräämiseen. Esimerkiksi paikallinen hakukone voi tarvita tietoa verkkosivun kielestä tehdessään päätöstä siitä, pitäisikö sivu indeksoida [MS05].

2 Automaattinen kielentunnistus teksteistä

Automaattinen kielentunnistus on yksi tiedon louhinnan osa-alue, sen tarkoitus on kertoa millä kielellä tai kielillä tietty teksti on kirjoitettu [ŘK09]. Kielentunnistus perustuu yleensä testiaineistosta luotuun kielen malliin, johon tuntemattomalla kielellä kirjoitettua tekstiä verrataan.

Tekstin metatiedot ovat yksi lähde kielentunnistuksessa [MS05]. Perinteisessä korpuksessa metatietoihin voi pääsääntöisesti luottaa ja ne tarjoavat usein kielen lisäksi laajempaakin tietoa alkuperäisistä teksteistä. Perinteisen korpuksen tekstien metatiedoista voi käydä ilmi kirjoittajan ikä, sukupuoli, tekstin tyyppi ja jopa se mihin käyttöön teksti on tarkoitettu. Näitä tietoja voidaan luonnollisesti hyödyntää analyysien teossa. Verkkoteksteissäkin saattaa olla metatietoja, mutta niihin luottaminen ei ole aivan yhtä helppoa.

Automaattinen kielentunnistus toimii siten, että testivaiheessa luodaan malli kielestä joihin tutkittavaa tekstiä sitten verrataan tunnistusvaiheessa. Kielentunnistus on luokitteluongelma siinä mielessä, että kohdekielet täytyy tietää etukäteen tekstin kieltä tunnistettaessa [ŘK09]. Verkkotekstejä tutkittaessa testiaineisto voidaan rakentaa esimerkiksi perinteisestä korpuksesta, jonka kieli on tiedossa. Tämän testiaineiston tarkoitus on siis luoda kielen malli, johon uusia tuntemattoman kielen verkkotekstejä verrataan.

Automaattisen kielentunnistuksen tehokkuuteen vaikuttaa suorasti testiaineiston koko ja annetun tekstin koko. Tekstit voidaan luokitella pieneksi (30 merkkiä tai korkeintaan viisi sanaa), isoksi (yli 300 merkkiä tai 50 sanaa) ja keskikokoiseksi (lause, ison ja pienen väliltä) [ŘK09].

2.1 N-grammit

N-grammit ovat merkkiyhdistelmistä muodostettuja taulukoita, joihin yleisesti käytössä olevat tehokkaat ja toimivat kielentunnistusmenetelmät perustuvat. On huomattu, että ihminen tarvitsee yllättävän vähän tunnistaakseen kielen, vaikka hän ei osaisikaan kieltä sujuvasti ja teksti olisi hyvin lyhyt [ŘK09]. Tästä voidaan päätellä, että erilaiset merkkien yhdistelmät antavat kielelle sen ominaiset piirteet ja siten merkkiperustaiset n-grammit ovat hyvä keino erottaa kielet toisistaan. Tiettyjen merkkien yhdistelmät luovat hyvän kuvan kielestä ja ovat ominaisia juuri tälle kielelle.

Kahden tai kolmen merkin yhdistelmien n-grammit on todettu toimivimmiksi, kolmen merkin n-grammit jopa tehokkaammiksi kuin kahden [SCH+94]. Kolmen merkin yhdistelmiä käytettäessä optimaalinen kielentunnistus on saavutettu jo kun läpikäyty 25-50% kielen testiaineiston merkkiyhdistelmistä, kun kahden merkin yhdistelmiä tarvitaan optimaaliseen suoritukseen 75% kielen testiaineiston merkkiyhdistelmistä.

Niin sanottu Markovin malli (*Markov model*) käyttää n-grammeja laskeakseen merkkiyhdistelmien esiintyvyyden todennäköisyyksiä niiden sekvenssien perusteella ja antaa tulokset sen perusteella, millä kielellä on suurin todennäköisyys tuottaa annettu teksti [ŘK09].

N-grammit ovat todella virhesietoisia, ja siksi soveltuvat hyvin käytettäväksi verkkotekstien kielentunnistukseen [MS05]. Kielen mallintammista n-grammeilla puolustaa myös menetelmän vakaus. Ongelmallisetkin kielen ilmiöt kuten uudet sanat, kielivirheet ja harvinaiset taivutusmuodot voidaan yhdistää kieleen n-grammien avulla.

Testiaineiston osalta enempi ei välttämättä ole parempi, jo suhteellisen pienellä joukolla saadaan näyttävä malli kielestä [ŘK09].

N-grammeja käytettäessä tekstin tyypillä on väliä. Annetun tekstin olisi oltava samaa tekstityyppiä kuin testiaineiston, jos halutaan luotettavia tuloksia [SCH+94]. Verkossa tekstin tyyppi ei aina ole selvillä.

2.2 Yleisten sanojen lista

Yleisten sanojen käyttö kielentunnistuksessa perustuu yleensä kielessä esiintyviin pieniin "vähämerkityksisiin"sanoihin, joita esiintyy suhteellisen paljon tämän kielen tekstissä [ŘK09]. Voidaan puhua myös pienten sanojen tekniikasta. Tätä menetelmää käytettäessä testiaineistolle on aluksi tehtävä lauseopillinen analyysi, jotta tarvittavat väli- ja loppusanat, prepositiot ja konjunktiot saataisiin listattua [FdSPL06].

Kielentunnistuksessa voidaan myös käyttää esimerkiksi kielen 100 yleisintä sanaa [SCH+94]. Testiaineistosta luodaan sanan esiintyvyyden mukaan järjestetty sanojen lista. Annettua tekstiä läpikäydään eri kielten yleisten sanojen listoja käyttäen, kasvatetaan kielen laskuria aina kun kohdataan sen kielen yleinen sana. Prosessin missä tahansa vaiheessa voidaan palauttaa todennäköisin kieli tähän mennessä. Kun koko teksti on käyty läpi, verrataan laskureita ja suurin arvo kertoo mistä kielestä on kyse.

Yleisten sanojen menetelmää on kritisoitu siitä, että se on liian rajoittunut ja soveltuu vain pidemmille teksteille [ŘK09].

2.3 Muita menetelmiä

Erityisiä, uniikisti johonkin kieleen kuuluvia merkkiyhdistelmiä käytetään liittämään teksti tähän kieleen [SCH+94]. Samoin kuin edellä kuvatuissa menetelmissä, automaattinen kielentunnistus käyttää testiaineistoa luodakseen kielen mallin, ja tähän malliin sitten verrataan annettua tekstiä. Menetelmä toimii kehnosti ainakin lyhyillä teksteillä, joissa suhteellisen harvinaisia uniikkeja merkkiyhdistelmiä ei välttämättä esiinny. Testiaineisto ei välttämättä sisällä yhtään kielelle uniikkeja merkkiyhdistelmiä, täten kyseistä kieltä olisi mahdoton tunnistaa pelkästään tätä menetelmää käyttämällä. Ongelmana on lisäksi se, että testiaineiston perusteella kaikkia kielen uniikkeja merkkiyhdistelmiä ei luonnollisestikaan voida listata.

Menetelmää käytettäessä myös tutkittavien kielten joukolla on merkitystä, uniikit merkkiyhdistelmät selvitetään vertaamalla eri kielten testiaineistoja toisiinsa. Jollain testiaineistojen joukolla merkkiyhdistelmä saattaa olla uniikki tietylle kielelle, jollain toisella joukolla kyseistä yhdistelmää voisi löytyä useammastakin kielestä. Näin tulokset eivät aina ole luotettavia, koska kieltä ei voida tunnistaa varmuudella. Tuhansien lauseiden testiaineistolla voitaisiin jo löytää tarpeeksi kieleen kuuluvia uniikkeja merkkiyhdistelmiä, mutta niin suurten testiaineistojen käsittely ei ole mielekästä.

Verkkosivun kieli voidaan joissain tapauksissa tunnistaa pelkän URL:in perusteella [BHW08]. Menetelmästä on se hyöty, että sivun tekstiä ei tarvitse turhaan ladata tutkittavaksi, jos se ei vastaa haluttua kieltä. Kielispesifit hakukoneet (*yandex.ru*, *fireball.de*) hyötyvät menetelmästä, koska ne haluavat koota tuloksiinsa vain tietyn kielisiä verkkosivuja. Domainin maakoodi yksistään ei kerro luotettavasti kohdesivun kieltä, ja monissa maissa on useampi kuin yksi virallinen kieli. Menetelmän kielentunnistus perustuu koneoppimisen algoritmeihin, joilla on käytössään erilaisia sanakirjoja (mm. kaupunkien nimistä). Suurena haasteena kielentunnistuksessa URL:in perusteella on englannilta näyttävien ei-englanninkielisten verkkosivujen URL:it. Ihminenkin suoriutuu tehtävästä joissain tilanteissa algoritmia huonommin.

3 WWW ja vaikeat tapaukset

World Wide Web on nostanut kielentunnistuksen yhä tärkeämpään asemaan. Verkkotekstit ovat kuitenkin haastellisempia tunnistettavia monestakin syystä. Kielivirheet ovat verkossa yleisiä, samoin kuin monikieliset tekstit. Lyhyidenkin tekstien kieli halutaan tunnistaa. Automaattista kielentunnistusta käytetään myös, kun paikallinen hakukone löytää vieraasta domainista tekstin, ja tutkii pitäisikö se indeksoida. Lähekkäiset kielet on vaikea erottaa toisistaan, toisaalta myös saksa ja englanti tunnistetaan ajoittain toisikseen johtuen niiden yhteisestä historiallisesta taustasta [MS05].

Verkossa julkaisukynnys on paljon alhaisempi, kuin painetussa tekstissä [KG03]. Tämä lisää kielivirheitä ja luo kielentunnistuksellekin haasteen. Automaattisen kielentunnistuksen on osattava tunnistaa osittain virheellinenkin kieli.

Automaattisen kielentunnistuksen todellinen haaste on hyvin lyhyissä teksteissä, monikielisissä teksteissä [ŘK09] ja toisiaan lähellä olevien kielten ja kielivarianttien erottamisessa toisistaan [LMB07]. Verkkotekstit on luotu ihmiskäyttäjää eli lukijaa ajatellen, ja siksi niiden automaattinen käsittely ja tulkinta on toisinaan vaikeaa [ŘK09]. XML ja semanttinen merkintä yrittävät tarjota ratkaisua tähän ongelmaan, mutta todellisuudessa monissa dokumenteissa on puuttuvat tai täysin virheelliset metatiedot [ŘK09]. Metatietoja ei täten voida käyttää luotettavana lähteenä kielentunnistuksessa.

Verkossa tekstin tyyppi ei ole aina selvillä, näin esimerkiksi n-grammeja käyttävät menetelmät eivät toimi optimaalisesti.

Eräs merkittävä haaste on myös tekstit, jotka on kirjoittettu tunnistamattomalla kielellä [ŘK09]. Yleisesti käytetyt menetelmät (n-grammit, yleisten sanojen lista) eivät tarjoa vaihtoehtoa "tuntematon kieli". Menetelmät lähtevät siitä, että kielten joukko on tunnettu ja niistä on olemassa testiaineistot, ja että annettu teksti on kirjoitettu jollakin näistä kielistä. Verkossa, kuin muussakaan kontekstissa, tämä ei aina toteudu. Ratkaisuna tuntemattoman kielen ongelmaan, voidaan käyttää vaatimusta että tekstin kielen täytyy ylittää määrätty minimiarvo testiaineistolla tuotettuun kielen malliin verrattaessa [ŘK09]. Tekstiä ei täten "väkisin"tunnistettaisi

siksi kieleksi, jota se parhaiten vastaa. Tämän minimiarvon asettaminen on kuitenkin osoittautunut ongelmalliseksi.

Mikään edellisissä luvuissa esitetyistä automaattisista kielentunnistusmenetelmistä ei sinällään toimi riittävän hyvin vaikeassa kontekstissa. N-grammeja voidaan hyödyntää, mutta verkkotekstien osalta on otettava huomioon tiettyjä seikkoja [MS05]. Ensiksikin verkkotekstistä on siivottava mielenkiinnottomat ja mahdolliset toistuvat osuudet pois, kuten valikot ja sähköpostiosoitteet. Lisäksi siivouksen avuksi voidaan ottaa Internetin yleisten sanojen ja fraasien lista, jolla karsitaan sivulta usein esiintyvät tarpeettomat fraasit kuten *java* ja *Made for Internet Explorer*. Metatietoihin voidaan päättää luottaa, ja esimerkiksi olettaa kieli metatiedoissa ilmoitetuksi kieleksi jos se on muu kuin englanti, ja täten ohittaa n-grammien ehdottama kieli. Näiden seikkojen lisäksi vielä monikieliset tekstin osat ja tunnistamaton kieli on osattava käsitellä, pelkällä n-grammien käytöllä ei tähän pystytä. Eräs tapa käyttää n-grammeja verkkosivun kielentunnistuksessa, on painottaa otsikon n-grammeja kolminkertaisesti ja kuvaavien metatagien n-grammeja kaksinkertaisesti [MS05].

Sanakirjamenetelmä [ŘK09] perustuu sanoihin merkkien sijaan. Yleisten sanojen sijaan käytetään sanoja, jotka ovat ominaisia juuri tietylle kielelle ja sitä kuinka ominaisia ne ovat. Menetelmä pohjautuu sanojen relevanssiin kielessä (*relevance mapping*). Relevanssi voi olla positiivinen, tällöin sanan olemassaolo viittaa kieleen. Vastaavasti negatiivisen relevanssin mukaan sanan poissaolo viittaa kieleen. Nollarelevanssi merkitsee, että korrelaatiota ei ole. Menetelmä vaatii suuremman testiaineiston kuin n-grammeja käyttävät menetelmät.

Kiellettyjen sanojen sääntöä voidaan käyttää hyödyksi [LMB07],mutta tällöinkin tiedossa täytyy olla mitä kieliä halutaan erottaa toisistaan, ja tuntematonta kieltä ei edelleenkään pystytä tunnistamaan.

3.1 Hyvin lyhyet tekstit

Mitä lyhyempi teksti on, sitä vähemmän automaattisella kielentunnistuksella on materiaalia verrattavana testiaineistoon. Lyhyillä teksteillä kielentunnistus on hankalaa.

Eräs menetelmä käyttää lyhyiden tekstien ja mahdollisesti monikielisten tekstien kielentunnistukseen merkkiyhdistelmistä laskettua diskriminantin neliötä (*quadratic dicriminant score*) [FdSPL06]. Merkkiyhdistelmä, joka erottaa kielen vahvasti muista kielistä, on vahva diskriminantti. Vastaavasti heikko diskriminantti on sellainen yhdistelmä, joka ei ole tyypillinen millekään kielelle. Testivaiheessa luodaan moniulotteinen tila, riippuen kielien määrästä ja niiden samanlaisuudesta. Luokitteluvaiheessa esitetään uudet tekstit vektoreina tässä tilassa ja luokitellaan ne käyttäen diskriminantin neliötä.

Hyvin lyhyissä teksteissä, esimerkiksi turisteille kohdistetuissa mainoksissa verkossa, ongelmaksi muodostuu myös paikallisen kielen paikannimet. Toisaalta mainos on kohdistettu matkailijoille, eli kielenä käytetään usein englantia. Kuitenkin esimer-

kiksi ravintolan ruokalistassa, ruokalajit paikallisella on kielellä mutta muu teksti englanniksi [FdSPL06].

Joissain tapauksissa voi olla tarkoituksenmukaista osoittaa liian pienet tekstit, esimerkiksi alle 40 merkkiä pitkät tekstit, tuntemattoman kielen luokkaan [MS05]. Tämän luokan tekstit käydään lopuksi manuaalisesti läpi, jotkut kielet jäävät edelleen tunnnistamatta, mutta osa pystytään osoittamaan oikeaan kieleen. Automaatti ei välttämättä tähän pystyisi, ja aiheuttaisi täten vääristymää aineistoon.

Sanarelevanssin käyttö on tuottanut kohtalaisia tuloksia myös lyhyillä teksteillä [ŘK09].

3.2 Monikieliset tekstit

Verkkoteksteissä monikielisyys on yleistä. Verkkosivun looginen rakenne voi olla syynä useamman kuin yhden kielen käyttöön. Esimerkiksi sivupalkki, tekijänoikeudet ja itse teksti voivat olla eri kielillä kirjoitettuja. Myös tekstin luonne voi vaatia useamman kielen käyttöä [ŘK09]. Tämä aiheuttaa sen, että dokumentti ei vastaakaan suoraan mitään kieltä ja kielen malli sekoittuu. Tuloksena voi olla, että kieli tunnistetaan joksikin aivan muuksi kuin mitä se on.

Eräs menetelmä monikielisten tekstien kielentunnistamiseen käyttää tekstin segmentointia ja yksittäisten segmenttien käsittelyä [ŘK09]. Kun teksti on luokiteltu yksikielisiksi blokeiksi, niitä voidaan käsitellä kuin yksikielistä tekstiä ja aineiston vääristyminen "väärän"kielen takia ei ole ongelma. Segmentointi on hyvin raskas operaatio, vaikka sen avulla on saatukin merkkitasolla loistavia tuloksia kielentunnistuksessa. Kun joustetaan tulosten täsmällisyydestä, voidaan saada tehokkaampi ja selkeämpi algoritmi.

Toinen lähestymistapa on yrittää löytää sivun "pääkieli"tunnistamalla sen isoimman yhtenäisen tekstiosan kieli ja painottaa sen n-grammeja [MS05]. Tämä metodi ei kuitenkaan erottele sivun muita kieliä, eikä siten sovellu korpuksen muodostamisessa käytettäväksi. Paikallisen hakukoneen indeksointiin se kuitenkin jotenkin soveltuu.

Vektorimenetelmää voidaan myös käyttää kielentunnistukseen monikielisistä teksteistä [Pra99]. Ulottuvuudet tilassa on määritelty testiaineistojen merkkiyhdistelmien, sanojen ja niiden molempien perusteella. Menetelmällä voidaan luotettavasti myös erottaa toisistaan kielivariantit, kuten kirjanorja ja uusnorja. Tässä tapauksessa on siis mielekästä luokitella norjan kieli kahteen luokkaan, jotta testattavat tekstit tunnistetaan oikein.

3.3 Toisiaan lähellä olevat kielet ja kielivariantit

Automaattinen kielentunnistus toimii jopa sataprosenttisesti tilanteessa, jossa teksti on kirjoitettu vain yhdellä kielellä ja kun tekstin koko on tarpeeksi suuri. Kielentunnistusmenelmät eivät kuitenkaan toimi, kun tarvitaan tieto millä kielen varianteista teksti on kirjoitettu, esimerkiksi erottelemaan toisistaan euroopan portugali

ja brasilian portugali tai brittienglanti ja amerikan englanti [FdSPL06].

Slaavilaiset kielet, joilla on usein osittain samoja kielioppisääntöjä ja myös samaa sanastoa, on vaikea erottaa toisistaan. Verkkoteksteissä on yleistä, että kielen painomerkit jätetään kirjoittamatta, tällöin menetetään tärkeää kielellistä informaatiota ja kielet sekoittuvat toisiinsa [ŘK09].

Kroaatin, serbian, slovenian ja slovakin kielten tunnistaminen, tai nimenomaan erottaminen toisistaan, on haastava tehtävä, koska kielet muistuttavat hyvin paljon toisiaan. Eräs menetelmä käyttää erikoismerkkejä ja uniikkeja merkkiyhdistelmiä erotellakseen nämä kielet toisistaan [LMB07]. Esimerkiksi tunnettu kielentunnistusalgoritmi TextCat sekoittaa usein läheiset kielet. TextCat käyttää kielentunnistukseen merkkiperustaisia n-grammeja, annetusta tekstistä luodaan merkkiyhdistelmien n-grammi jota verrataan saatavilla oleviin malleihin, ja paras osuma voittaa. Mitä suurempi annettu teksti, sitä luotettavammin kieli voidaan tunnistaa [ŘK09].

Kiellettyjen sanojen lista auttaa tunnistuksessa ja kielten erotuksessa toisistaan [LMB07]. Kroatian ja serbian kielten tunnistamisessa on käytetty apuna listaa sanoista, jotka esiintyvät toisen testiaineistossa viisi tai enemmän kertaa ja toisen kielen testiaineistossa ei lainkaan. Näin siis voidaan erottaa kaksi kieltä toisistaan, mutta kontekstissa jossa tunnistettavia kieliä on useampia, menetelmää täytyisi kehittää.

Dokumenttien samanlaisuuden vertailua käytetään myös hyväksi kielentunnistuksessa. Menetelmällä on erotettu luotettavin tuloksin Euroopan portugali ja Brasilian portugali toisistaan. Tutkimuksessa käytettiin dokumenttien samanlaisuuden vertailua n-grammien avulla ja vahvoja diskriminantteja [FdSPL06].

Portugalilaisen hakukoneen indeksointia varten tarvitaan automaattista kielentunnistusta, mutta se yksin ei riitä. Hakukone ei halua listata brasilialaisia sivuja, joten kielivariantit on kyettävä erottamaan toisistaan. Tietyt erikoiset merkit esiintyvät vain brasilian variantissa, näiden tunnistaminen voisi auttaa luokittelussa. Apumenetelmäksi on myös ehdotettu hyperlinkkien analysointia, koska hyvin usein esimerkiksi brasilialaiselta sivulta on linkkejä brasilialaisille sivuille. Whois-tietokantojen hyödyntäminen voisi antaa myös lisätietoa maasta, koska tietueet sisältävät domainien vastuuhenkilöiden yhteystiedot [MS05].

4 Edustavuudesta

Voidaan kysyä onko verkko korpus, jolloin päästään korpuksen määritelmän äärelle: Mikä on korpus, ja mikä on hyvä korpus? Korpuksen löyhimmän määritelmän mukaan se on vain joukko tekstejä [KG03], tämän määritelmän mukaan verkko on korpus, tarkemmin ottaen monikielinen korpus. Verkkotekstien käyttö korpuksen muodostuksessa tuo edustavuuteen uuden ulottuvuuden.

Tekstityyppi vaikuttaa korpuksen edustavuuteen, verkkoteksteistä koottu korpus kuvaa vain oman tyyppinsä tekstejä. Verkko edustaa vain itseään, kuten muutkin

korpukset. Ollakseen edustava, korpuksen on oltava tasapainotettu. Jos korpuksesta halutaan tehdä laajempia yleistyksiä, sen täytyy edustaa laajasti tekstityyppejä. Esimerkiksi keskustelufoorumeilta aineistoa kerättäessä täytyy muistaa, että joukko kuvaa keskustelufoorumeilla käytettyä kieltä ja sen laajemmissa yleistyksissä täytyy olla varovainen.

Alhainen julkaisukynnys verkossa vaikuttaa tekstien kielivirheisiin, verkkoteksteissä on enemmän virheitä kuin painetussa tekstissä. Esimerkki googletus virheellisellä sanalla tuo monta osumaa [KG03].

Korpuksen on oltava edustava tekstityypeiltään, jos siitä halutaan tehdä laajempia yleistyksiä [Bib93]. Verkkoteksteissä tekstin tyyppi ei ole aina tiedossa, lisäksi joukko voi olla dynaaminen (esimerkiksi keskustelufoorumit). Tilanteessa jossa haetaan esimerkiksi tekstejä tietyltä uutissivulta, voidaan haluta vain uutistekstejä. Nykyisillä uutissivustoilla on kuitenkin paljon muutakin, esimerkiksi lukijoiden kommentit uutisista. Kommentit voivat lisäksi olla eri kielillä kuin uutinen on kirjoitettu. Samaan joukkoon kuuluu blogitekstit, ja lukijoiden kommentit blogiteksteihin. Verkkosivun looginen rakenne tosin tässä tapauksessa kertoo, mitä tekstiä mikin osa edustaa. Uutistekstit voidaan erottaa kommenttiteksteistä ja täten valita tutkimuskohteeksi vain jompikumpi.

Otoksen koko on tärkeä tekijä edustavuuden näkökulmasta. Tärkeämpää kuitenkin on otosten valinnan suunnittelu ja perusjoukon määrittely [Bib93]. Edustavuus kertoo missä mittakaavassa otoksesta voidaan tehdä yleistyksiä koko perusjoukkoon. Edustavan korpuksen takaamiseksi yksi menetelmä on tehdä prosessista syklinen. Ensin tehdään teoreettiseen kehykseen ja pilottitutkimuksen analyysiin perustuva suunnitelma, jonka mukaan kerätään korpus, arvioidaan saatua aineistoa ja korjataan suunnitelmaa. Tätä kautta saadaan edustavampi korpus.

Pilottikorpuksen käyttö toimisi verkkotekstienkin kohdalla, vaikka joitakin näkökohtia on otettava huomioon. Jos verrataan verkosta kerättyä korpusta perinteiseen korpukseen, erona on joukon dynaamisuus. Kuvitellaan että perinteinen korpus on kerätty suomenkielisestä 1990-luvun kaunokirjallisuudesta, silloin sen joukko on olemassaoleva ja muuttumaton. Verkko sen sijaan on jatkuvasti muuttuva ja kasvava joukko tekstejä. Jos korpus kerätään esimerkiksi tietyltä keskustelufoorumilta, joukon senhetkistä tilaa ei välttämättä kyetä enää tuottamaan uudelleen. On otettava käyttöön lisämääreitä, esimerkiksi käydyt keskustelut tiettyyn päivämäärään mennessä. Tämäkään ei takaa muuttumatonta joukkoa, koska yleensä keskustelufoorumin ominaisuuksiin kuuluu, että käyttäjä voi jälkikäteen muokata tai jopa poistaa kirjoittamiaan viestejä.

Edustavuutta ajatellen, on tietenkin oltava mielessä jo korpusta kerättäessä se, mitä halutaan tutkia ja kuvata. Jos tutkimuksen kohteena on nykysuomi, verkon suomenkielinen keskustelufoorumi voi olla yksi varteenotettava tutkimuskohde. Pelkästään se ei kuitenkaan riitä, vaikka sen kieli kuvaakin varmasti nykysuomea. Keskustelufoorumilla käytetty kieli ei kuvaa kieltä kattavasti ja edustavasti koska joukko on ainoastaan satunnaisten käyttäjien kirjoittamaa kieltä verkossa. Jos sen sijaan halutaan kuvata suomalaisen tietyn keskustelufoorumin kieltä, joukko on varmas-

ti hyvin edustava. Voidaanko korpuksesta sitten tehdä yleistyksiä suomen kieleen ja halutaanko niitä edes tehdä, on toinen kysymys. Toisaalta voidaan kysyä mitä järkeä on tutkimuksessa, joka ei mahdollista minkään tason yleisiä johtopäätökisä.

Täydellä varmuudella voidaan ennustaa vain perusjoukon tekstityypin kielenkäyttöä, ja oikeastaan tämäkin vain jos testiaineisto ja kerätyt tekstit ovat saman perusjoukon satunnaisia otoksia [KG03]. Kielen malleja luotaessa verkkoteksteistä ongelma on seuraava: Jos halutaan luoda myös testiaineistot verkkoteksteistä, ensin on joka tapauksessa kerättävä ne ja tunnistettava niiden kieli. Tässä vaiheessa tarvitaan siis jo testiaineistoja, joiden avulla kieli tunnistetaan. On käytettävä jo olemassaolevia korpuksia tai muita aineistoja kielen mallin luomiseen. Tässä taas palataan kysymykseen tekstin tyypistä, kuten sanottua, tekstin tyyppi vaikuttaa kielen malliin. Verkkotekstien kielentunnistuksessa tulokset voivat vääristyä juuri tämän takia, kun kielen malli on luotu eri tekstityypin aineistosta kuin haetut tekstit.

Verkkoteksteistä kerätyssä korpuksessa haasteena on saada edustava otos juuri siitä kielestä mitä halutaan kuvata. Yksi yleisesti käytettyjen menetelmien ongelma on tuntemattoman kielen käsittelyn puuttuminen. Kieli tunnistetaan väkisin joksikin kieleksi, mitä se eniten muistuttaa, vaikka se ei kuuluisi ollenkaan testiaineiston kielten joukkoon. Tuntemattoman kielen "tunnistaminen"parantaisi kielentunnistusta jättäen tunnistamattomat kielet aineistosta pois, ja täten edustavuus paranisi.

Edustavuus kärsii myös kun tunnistetaan vain osatekstin kieli, näin käy usein monikielisten verkkotekstien tapauksessa. Segmentointi ratkaisee pulman osittain, mutta on raskas operaatio. Menetelmän täytyy tekstin paloittelun lisäksi vielä tietää miten erikieliset blokit sitten käsitellään. Menetelmän täytyy myös osata jättää tuntematon kieli luokittelematta, jotta aineisto ei vääristy.

Verkkoteksteistä tuotetuissa kielen malleissa tiettyjen sanojen ja fraasien yliedustavuus voi vääristää aineistoa [ŘK09]. Nämä Internetin yleiset fraasit ovat yleensä englantia, mutta ne eivät välttämättä kerro siitä että sivun kieli on englanti. Automaatin täytyy osata käsitellä näitä fraaseja. Kun verkkosivu on ladattu tutkittavaksi, sitä täytyy siivota tarpeettomasta materiaalista. Tässä siivouksessa halutaan päästä eroon esimerkiksi HTML-tageista, ja samalla voidaan tiputtaa aineistosta pois myös nämä edellä mainitut yleiset sanat ja fraasit.

5 Yhteenveto

Automaattinen kielentunnistus on alue, jota on tutkittu paljon. On löydetty tehokkaita menetelmiä, jotka toimivat loistavasti tietyissä tilanteissa. Vaikeassa kontekstissa nämä yleiset menetelmät eivät kuitenkaan toimi tarpeeksi hyvin, aiheuttaen sen että kieli tunnistetaan joksikin muuksi kuin itsekseen tai vain osa tekstin kielestä tunnistetaan. Saatu aineisto ei näin ole hyvää tutkimusmateraalia, koska se ei edusta varsinaisesti mitään kieltä.

Toimivampia ratkaisuja on ehdotettu vaikeisiin tilanteisiin. Menetelmiä on enemmänkin kuin tässä raportissa on esitelty: kielentunnistuksessa on hyödynnetty myös

esimerkiksi suhteellista entropiaa sekä kielen morfologiaa ja syntaksia [LMB07].

Automaattinen kielentunnistus ei ole vielä loppuunkoluttu tutkimusalue. Tarvitaan yleisesti mihin tahansa kontekstiin ja mille tahansa kielille toimiva menetelmä, ei kieli- ja ympäristöspesifiä tekniikkaa. Verkkotekstejä kerättäessä korpukseksi ei vielä tiedetä onko kohdesivu esimerkiksi mahdollisesti monikielinen, siksi ei ole mielekästä määritellä vain tiettyyn tilanteeseen sopivaa menetelmää. Tunnistettavien kielten joukko on oltava etukäteen tiedossa, mutta myös tuntemattoman kielen tunnistus parantaa luokittelua.

Tietyt vaikeat tapaukset tulevat kuitenkin mitä todennäköisimmin pysymään vaikeina. Esimerkiksi kahden hyvin samankaltaisen kielen tai kielivariantin erottaminen on vaikea tehtävä ihmisellekin, joten se tulee väistämättä olemaan vaikea kielentunnistusautomaatillekin.

Lähteet

BHW08 Baykan, E., Henzinger, M. ja Weber, I., Web page language identification based on urls. *Proc. VLDB Endow.*, 1,1(2008), sivut 176–187.

Bib93 Biber, D., Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8,4(1993), sivut 243–257. URL `http://llc.oxfordjournals.org/content/8/4/243.abstract`.

FdSPL06 Ferreira da Silva, J. ja Pereira Lopes, G., Identification of document language is not yet a completely solved problem. *CIMCA '06: Proceedings of the International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce*, Washington, DC, USA, 2006, IEEE Computer Society, sivu 212.

KG03 Kilgarriff, A. ja Grefenstette, G., Introduction to the special issue on the web as corpus. *Comput. Linguist.*, 29,3(2003), sivut 333–347.

LMB07 Ljubesic, N., Mikelic, N. ja Boras, D., Language indentification: How to distinguish similar languages? *Information Technology Interfaces, 2007. ITI 2007. 29th International Conference on*, 2007, sivut 541 –546.

MS05 Martins, B. ja Silva, M. J., Language identification in web pages. *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, New York, NY, USA, 2005, ACM, sivut 764–768.

Pra99 Prager, J. M., Linguini: Language identification for multilingual documents. *HICSS '99: Proceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences-Volume 2*, Washington, DC, USA, 1999, IEEE Computer Society, sivu 2035.

ŘK09 Řehůřek, R. ja Kolkus, M., Language identification on the web: Extending the dictionary method. Teoksessa *Computational Linguistics and Intelligent Text Processing*, Gelbukh, A., toimittaja, osa 5449 sarjasta *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2009, sivut 357–368, URL `http://dx.doi.org/10.1007/978-3-642-00382-0_29`.

SCH+94 Souter, C., Churcher, G., Hayes, J., Hughes, J. ja Johnson, S., Natural Language Identification using Corpus- Based Models. *Hermes, Journal of Linguistics*, 1,13(1994), sivut 183–204. URL `http://download2.hermes.asb.dk/archive/FreeH/H13_15.pdf`.