Suomenkielinen puhehaku

Inger Ekman

Tampereen yliopisto
Tietojenkäsittelytieteiden laitos
Pro gradu -tutkielma,
tietojenkäsittelyoppi
Huhtikuu 2003

Tampereen yliopisto
Tietojenkäsittelytieteiden laitos
Inger Ekman: Suomenkielinen puhehaku
Pro gradu -tutkielma, 74 sivua
Huhtikuu 2003

Tiedonhakutekniikoilla pyritään tarjoamaan ihmisille mahdollisuus hakea tietokantojen sisältämiä dokumentteja sisältöperusteisesti vapaamuotoisilla kyselyillä. Puhetiedonhaun tavoitteena on saattaa myös puheena tallennettu tieto ihmisten käytettäväksi.

Tiedonhakumenetelmien kehittämisessä on keskitytty erityisesti tekstimuotoon tallennetun tiedon hakemiseen. Audiomuotoisiin tallenteisiin ei suoraan voi käyttää perinteisiä tekstitiedonhaun menetelmiä. Siksi tarvitaan uusia tiedonhakumenetelmiä, jotka soveltuvat puhehakuun. Tässä työssä käsitellään puhemateriaaliin kohdistuvia tiedonhakumenetelmiä. Erityisesti tarkastellaan, miten puhehaussa käytetyt menetelmät soveltuvat suomenkielisen puhemateriaalin hakemiseen.

Työn kokeellisessa osuudessa rakennettiin suomenkielisen puhehakujärjestelmän prototyyppi, jolla tutkitaan n-grammien avulla suoritettavan suodatuksen soveltuvuutta suomenkieliseen puhetiedonhakuun. N-grammit ovat menetelmäperhe, jossa dokumenttien samankaltaisuutta verrataan niiden sisältämien n merkin mittaisten merkkijonojen perusteella. Käyttämällä n-grammeja yhdessä nimikirjoitusten kanssa voidaan tehokkaasti käsitellä isoja datamääriä

Tutkimuksessa verrataan eri n arvojen vaikutusta n-grammien avulla muodostettavien nimikirjoitusten suodatuskykyyn. Suodatuksessa käytettäviä nimikirjoituksia muodostetaan sekä kokonaisista puhedokumenteista että puhedokumenttien osista. Suodatus suoritetaan yksittäisten hakusanojen perusteella, mutta menetelmää voi helposti laajentaa kokonaisten kyselyjen käsittelyyn. Kokeiden testiaineistona on käytetty suomenkielistä uutismateriaalia. Tutkimuksessa verrataan suodatusmenetelmien toimintaa sekä puhuttuja että kirjoitettuja hakusanoja käyttämällä.

Avainsanat ja -sanonnat: Puhehaku, puhetiedonhaku, tiedonhaku, suomen kieli, n-gram, osittaistäsmäytys, äännetunnistus.

Sisällysluettelo

1 Johdanto

Tiedonhakutekniikoilla pyritään tarjoamaan ihmisille mahdollisuus hakea tietokantojen sisältämiä dokumentteja sisältöperusteisesti vapaamuotoisilla kyselyillä. Tiedonhakumenetelmien kehittämisessä on keskitytty erityisesti tekstimuotoon tallennetun tiedon hakemiseen. Puhetiedonhaun tavoitteena on saattaa myös puheena tallennettu tieto ihmisten käytettäväksi.

Puhuttua tietoa tuotetaan valtavia määriä päivittäin eri tarkoituksiin. Radio-ohjelmat ovat eräs esimerkki puhutusta tiedosta. Tilastokeskuksen mukaan sekä yksityisten radioasemien että Yleisradion kanavien lähetysaika on ollut vuodesta 1995 lähtien pääsääntöisesti 24 tuntia vuorokaudessa kanavaa kohden. (Tilastokeskus 2003a; 2003b) Tämä tarkoittaa sitä, että jo pelkästään Suomen radioasemissa päivittäin tuotetaan yhteensä useita tuhansia tunteja audiomateriaalia, josta suurin osa sisältää puheetta. Audiotallenteiden lisäksi puheella välitetään usein paljon informaatiota esimerkiksi televisio-ohjelmissa. Myös verkko-opetuksen yhteydessä on korostettu puhutun tiedon merkitystä (Leppävirta 2001). Esimerkkinä opiskelijoille hyödyllisestä materiaalista käy vaikkapa nauhoitus etäopetuksena järjestetystä luennosta. Puhetta käytetään paljon myös tieteellisessä tutkimuksessa. Esimerkiksi haastattelunauhat sisältävät puhetta, joiden sisältöperusteinen hakeminen nopeuttaisi aineiston käsittelyä merkittävästi.

Toisin kuin teksti, puhe on aikasidonnainen tapa välittää informaatiota. Puhetta ei esimerkiksi voi silmäillä läpi niin kuin kirjaa, vaan puheen ymmärtäminen edellyttää, että sitä kuunnellaan likimain samalla nopeudella kuin se on tuotettu. Kuten edellä esitin, puhemateriaalia on paljon ja sen läpikäyminen kuuntelemalla kestää kauan. Siksi on erityisen tärkeää kehittää sellaisia hakujärjestelmiä, joilla puheella esitettyä tietoa voi hakea.

Tekstitiedonhaun menetelmiä voidaan käyttää puheen lähdemateriaalina olleiden tekstien hakemiseen. Ongelma on, että osalla puhumalla tuotetusta materiaalista ei ole olemassa minkäänlaista tekstimuotoista esitystä. Audiomuotoisiin tallenteisiin ei myöskään suoraan voi käyttää perinteisiä tekstitiedonhaun menetelmiä. Jotta audio- ja multimediatiedostojen sisältämiä valtavia tietomääriä pystytään hakemaan, tarvitaan uusia tiedonhakumenetelmiä, jotka soveltuvat puhehakuun. Puhehaun tavoitteena on käsitellä audiomuotoista tietoa automaattisen

tietojenkäsittelyn keinoin siten, että tiedonhaku kohdistetaan audiotallenteisiin. Toisin sanoen niitä ei ihmisen toimesta muokata millään tarvoin ennen tiedonhakua.

Tämä tutkielma liittyy osana USIX/3-hankkeen projektiin Suomenkielisen puhemateriaalin haku. Projekti oli kolmivuotinen Tampereen teknillisen yliopiston signaalinkäsittelyn laitoksen ja Tampereen yliopiston informaatiotutkimuksen laitoksen yhteisprojekti, jossa työskentelin osa-aikaisena tutkijana vuosina 2001-2002.

Tämän työn tarkoitus on esitellä menetelmiä puheesta suoritettavalle tiedonhaulle. Erityisesti pohdin, miten puhehaussa käytetyt menetelmät soveltuvat suomenkielisen puhemateriaalin hakemiseen.

Työn kokeellisessa osuudessa rakensin suomenkielisen puhehakujärjestelmän prototyypin. Järjestelmällä tutkin n-grammien avulla suoritettavan suodatuksen soveltuvuutta suomenkieliseen puhetiedonhakuun. Suodatuksessa tavoitteena on löytää tietokannasta joukko dokumentteja, joiden joukossa hakupyynnön kannalta tärkeät dokumentit todennäköisesti ovat. Kun tällainen joukko dokumentteja on löydetty, varsinaiset hakumenetelmät voidaan kohdistaa tähän joukkoon.

N-grammit ovat menetelmäperhe, jossa dokumenttien samankaltaisuutta verrataan niiden sisältämien *n* merkin mittaisten merkkijonojen perusteella (n-grammien käyttöä puhehaussa käsitellään luvussa 5.3). Käyttämällä n-grammeja yhdessä nimikirjoitusten kanssa voidaan tehokkaasti käsitellä isoja datamääriä (ks. esim. Robertson & Willett 1998). Nimikirjoitustiedostot ovat *hajatauluihin* (hash table) perustuvia tiedostoja, joissa yksittäisillä biteillä pyritään kuvaamaan kokonaisten dokumenttien tai niiden osien informaatiosisältöä. N-grammien tapauksessa tiedostot muodostetaan siten, että grammattava dokumentti tai merkkijono ensin jaetaan n-mittaisiin, osittain päällekkäisiin, merkkijonoihin. Tämän jälkeen dokumentin bittivektorissa kyseistä merkkiparia, -triplettiä tai -sekvenssiä kohden vastaava bitti muutetaan nollasta yhdeksi. (Ks. esim Järvelin 1995, 131–133; Ashford & Willett 1988, 89–91)

Tässä tutkimuksessa vertaan eri *n* arvojen vaikutusta n-grammien avulla muodostettavien nimikirjoitusten suodatuskykyyn. Kokeiden testiaineistona käytetään suomenkielistä uutismateriaalia. Suodatuksessa käytettäviä nimikirjoituksia muodostetaan sekä kokonaisista puhedokumenteista että puhedokumenttien osista. Tässä työssä suodatus suoritetaan yksittäisten hakusanojen perusteella, mutta menetelmää voi helposti laajentaa kokonaisten kyselyjen käsittelyyn. Tutkimuksessa

verrataan suodatusmenetelmien toimintaa sekä puhuttuja että kirjoitettuja hakusanoja käyttämällä.

Tutkielman rakenne on seuraava: Luvussa 2 käsittelen puheen luonnetta ja puhetta tiedon tallennusmuotona. Luvussa 3 käyn läpi puheentunnistuksen prosessia, sekä olemassa olevien järjestelmien ominaisuuksia. Luku 4 keskittyy suomenkielisen tiedonhaun erityisongelmiin ja pohtii, millaisia erityisominaisuuksia suomen kieli asettaa puhemuotoisen informaation haun kehittämiselle. Luvussa 5 kartoitetaan puhehaussa käytettyjä menetelmiä. Luvussa 6 käsitellään puhedokumenttien suodattamista n-grammien avulla muodostettujen nimikirjoitusten avulla. Suodatuksen tuloksia arvioidaan alaluvussa 6.6. Lopuksi luvussa 7 pohditaan suomenkielisen puhehaun tilannetta ja kokeellisesta tutkimuksesta saatujen tulosten merkitystä suomenkielisen puhehaun kehitykselle.

2 Puhe

Audiotallenteena esiintyvän tiedon hakemisen ymmärtäminen edellyttää puheen peruselementtien tuntemista. Tässä luvussa esittelen lyhyesti äänikäyrän ominaisuuksia, äänen digitaalista taltiointia sekä vertailen tietokoneen ja toisaalta ihmisen tapaa vastaanottaa ja käsitellä puhesignaalia. Lisäksi käsittelen puheen tehtävää tiedonvälityksessä ihmistenvälisessä viestinnässä.

2.1 Puhe ääniaaltona

Puhe on ilman hiukkasten välittämiä paine-eroja. Paine-erot ovat ääniaaltoja jotka kulkevat ilmassa ilman hiukkasten paikallisesti tihentyessä ja ohentuessa. Yksittäistä aaltoa voidaan kuvata kertomalla sen taajuus, eli värähtelyjen määrä sekunnissa ja amplitudi, eli värähtelyjen laajuus. Taajuuden yksikkö on hertsi (Hz = 1/s). Amplitudi ilmoitetaan desibeleinä (dB). Äänen frekvenssi vaikuttaa kuulohavainnossamme lähinnä korkeuteen ja sen amplitudi äänen voimakkuuteen. Niin sanottu siniääni on sinikäyrän muotoinen ääni kaksiulotteisessa taajuus–amplitudi-avaruudessa. Siniääni kuulostaa lähinnä vihellykseltä. (Ks. esim. Weinschenck & Barker 2000, 46–50; Lyons 1998.)

Kun ilmassa kulkee samaan aikaan useita ääniaaltoja, ilman hiukkaset tihentyvät ja ohentuvat samaan aikaan usean ääniaallon seurauksena. Tällöin eri äänien aallot interferoivat, eli yhdistyvät yhdeksi aalloksi. Puhuessaan ihminen tuottaa äänihuulissaan kurkunpää-ääntä. Kurkunpää-ääni on kompleksinen ääniaalto joka koostuu useista siniäänistä. Niitä siniääniä, joista kurkunpää-äänen tapainen kompleksinen ääni muodostuu, nimitetään osasäveliksi (Wiik 1981, 129).

Kurkunpää-ääni ei kuitenkaan milloinkaan pääse sellaisenaan ulkoilmaan. Tämä johtuu siitä, että ääniväylän ontelot aiheuttavat ääneen resonanssia. Resonanssi syntyy, kun ääni törmää seinään aaltokäyrän paineen vaihtelun maksimikohdassa. Tällöin ääniaalto saa ”lisävauhtia”, jonka seurauksena sen amplitudi kasvaa. Jos ääniaalto törmää seinämään jossakin muussa aaltokäyrän vaiheessa, se vaimenee. Kun kurkunpää-ääni kulkee ihmisen puhe-elimistön läpi, tietyt taajuudet voimistuvat samalla kuin toiset heikkenevät. Ääniväylän resonanssiominaisuudet riippuvat siitä, minkä muotoinen ääniväylä on, eli kenen suuontelo on kyseessä ja millaisessa

asennossa se on. Muuttamalla suuontelon muotoa ja esimerkiksi kielen asentoa ihminen voi puheen aikana muuttaa ääniväylänsä resonanssiominaisuuksia ja sitä myöten ulostulevaa ääntä. (Mts. 123–139.)

2.2 Ihmisen kuulohavainto

Ihmisen korva jaetaan tavallisesti kolmeen tarkasteltavaan osaan: ulko-, väli- ja sisäkorvaan. Ulkokorvan korvakäytävä päättyy tärykalvoon. Korvakäytävään päätyvät ääniaallot pistävät tärykalvon värähtelemään. Tärykalvon sisäpuolella, välikorvassa, on kolme toisiinsa kiinnittynyttä luuta: vasara, alasin ja jalustin. Nämä luut välittävät ja voimistavat ääntä, mutta estävät myös kovin laajojen tärykalvon liikkeiden välittymisen sisäkorvaan. Sisäkorvan tärkein osa on simpukka, simpukan muotoinen ontelo, jossa ääniaaltoon perustuva mekaaninen liike muuttuu hermoimpulsseiksi (Nienstedt ym. 1992, 492–496).

Ihmiskorva pystyy käsittelemään taajuuksia, jotka ovat välillä 20Hz–20kHz. Jos äänen perusvärähtely laskee alle 20Hz:n, ihminen ei enää tajua ääntä yhtäjaksoisena, vaan erillisinä toisiaan seuraavina paukauksina. Voimakkuuden tajuamiseen vaikuttaa ratkaisevimmin äänen intensiteetti. Äänen intensiteetti ei ole kuitenkaan ainoa voimakkuuden tajuamiseen vaikuttava ominaisuus, sillä äänen frekvenssilläkin on vaikutus siihen, miten voimakkaana ääni havaitaan. Herkemmin ihminen kuulee sellaiset äänet, joiden frekvenssi on noin 1–5 kHz. Siksi esimeriksi 100Hz 70dB ääni kuullaan yhtä voimakkaana kuin 3000Hz 58dB. Foni on tajutun voimakkuuden yksikkö, kun desibeli on fysikaalinen intensiteetin yksikkö. Ihmisen kuulokynnys ilmoitetaan yleensä 0 foniksi. Desibeleissä tämä raja vaihtelee noin 0 ja 60 dB välillä. Korvan kipukynnys on noin 120 fonia (~110–125dB). Jos äänen intensiteetti on tätä voimakkaampi, ihminen tuntee korvassaan kipua eikä tajua värähtelyä enää äänenä. (Mts. 496–497; Wiik 1981, 154–156)

2.3 Digitaalisesti taltioitu puhe

Puheen taltioinnissa mikrofoni rekisteröi ilmasta ihmisen puhe-elimistön tuottamat ilman paine-erot. Vaikka puhe on *jatkuvaa* ja *analogista* (toisin sanoen ääni on jatkuvaa sekä ajan että voimakkuuden suhteen), se voidaan esittää *diskreetissä* ja *digitaalisessa* muodossa (eli numeerisesti voimakkuuksina eri ajanhetkinä). Audiotallenteen *näytteenottotaajuus* (sampling rate) ilmaisee kuinka monta näytettä äänestä on otettu

joka sekunti. Äänen *bittikoko* (bit rate) puolestaan ilmaisee kuinka monta bittiä jokaisen yksittäisen näytteenottopisteen voimakkuuden esittämiseen on varattu. (Ks. Lyons 1998, 1–49.)

Edellä kuvatuin menetelmin voidaan esittää sarja arvoja, jotka ovat approksimaatio alkuperäisestä puhutusta äänestä. Diskreetti signaali kuvaa ilmasta taltioitua painevaihtelua näytteenottotaajuuden tarkkuudella. Haluttaessa painevaihtelu voidaan tästä kuvauksesta tuottaa uudelleen kaiuttimien avulla, eli ääni voidaan toistaa. Puheäänen käsittelyn kannalta eräs tärkeimpiä tehtäviä on selvittää, minkälaisista siniäänistä puheäänessä esiintyvät kompleksiset äänet muodostuvat. Koska suuontelon muodonmuutokset vaikuttavat ulostulevan äänen osasävelien voimakkuuteen, osasävelien voimakkuuksia tutkimalla voidaan tehdä oletuksia suuontelon muodosta äänen tuottohetkellä. *Fourier-analyysin* avulla voidaan selvittää, mistä samanaikaisista siniäänistä puheen kompleksisen ääni muodostuu (ks. Lyons 1998, 49–128). Tämän jälkeen voidaan esimerkiksi muodostaa äänen *amplitudi-frekvenssi-spektri*, joka kuvaa äänen taajuusjakaumaa yksittäisenä ajan hetkenä.

Spektrissä olevien taajuushuippujen sijainnit ovat ihmisen havaitsemille äänille ainutlaatuisia. On todennäköistä, että ihmisaivoissa tapahtuu jotakin Fourier-analyysin kaltaista, kun me kuuntelemme erilaisia kompleksisia ääniä. (Wiik 1981, 129).

Äänen näytteenottotaajuus on yhteydessä Fourier-analyysin avulla saataviin taajuuksiin. Matemaatikko Harry Nyquist on esittänyt todistuksen Nyquist-taajuudesta. Äänitallenteen Nyquist-taajuus on puolet sen taltioinnissa käytetystä näytteenottotaajuudesta ja kaikki sitä korkeammat taajuudet menevät digitaalisessa taltioinnissa hukkaan. Jos esimerkiksi puhetta tallennetaan 16kHz taajuudella, Nyquist-taajuus on 8kHz, mikä tarkoittaa että kaikki yli 8kHz taajuudet kadotetaan tallenteesta. Ylemmät taajuudet nimittäin ”naamioituvat” alemmiksi taajuuksiksi, koska niiden aallot mahtuvat (osittain) näytteenottopisteiden väliin. Tätä kutsutaan *laskostumiseksi* (aliasing). (Ks. Lyons 1998, 26–32.)

Edellä todettiin, että ihmiskorva pystyy kuulemaan vain alle 20kHz taajuudet. Audiotallenteiden standardit näytteenottotaajuudet ovat 44.1kHz (CD) tai 48kHz (DAT). Näitä korkeammista tallennustaajuuksista ei ole vaikutusta ihmiskorvalla havaittuun äänenlaatuun. Puheen tallentamisessa voidaan lisäksi ottaa huomioon, että puheen kantama ääni-informaatio normaalisti mahtuu 0-5kHz välille, jolloin näytteenottotaajuudeksi riittäisi peräti 10kHz. Jopa puhelimissa käytettävä 8kHz

näytteenottotaajuus riittää puheen ymmärrettävyyden kannalta, mutta silloin tallenteen äänenlaatu on huono. Lisäksi useat äänet, esimerkiksi frikatiivit (”suhisevat” konsonantit, esim. /s/), tuottavat myös yli 5kHz taajuuksia, joiden tallennuksesta voi olla hyötyä puheentunnistuksessa. Tämän takia puheentunnistukseen suositellaan 16kHz näytteenottotaajuutta. Tällöin ylimmät tutkittavat taajuudet ovat 8kHz. (Robinson 1998.)

2.4 Puhe informaation välityksessä

Puheen avulla ihmiset välittävät tietoa toisilleen. Välityksen toimivuus perustuu siihen, että tiettyyn kieliryhmään kuuluvilla ihmisillä on yhteinen näkemys siitä, miten tiettyjä puheääniä eli *äänteitä* (phone, suom. myös fooni) pitää tulkita. Puhutun tiedon välitys perustuu puheäänteiden *distinktiivisisiin eroihin*, joiden perusteella kuulija luokittelee kaksi äännettä joko erilaisiksi tai samanlaisiksi niiden kielellisen tehtävän mukaan. Tällaista puhutun kielen pienintä merkitystä erottelevaa yksikköä kutsutaan *foneemiksi* (phoneme). Tässä työssä käytän foneemeista merkintätapaa /x/, jossa x on käsiteltävän kielen foneemi. Äänteistä on tapana käyttää merkintää [x], jossa x on tarkasteltavassa puheessa tiettynä ajan hetkenä tuotettu äänne. Lisäksi käytän pitkien äänteiden merkitsemiseen kahta peräkkäistä saman äänteen merkkiä, toisin sanoen pitkä [x] olisi [xx]. (Vrt. IPA 1999.)

Äänteiden jako foneemeihin vaihtelee kielijärjestelmittäin. Esimerkiksi suomenkielessä äänteiden [i] ja [u] välillä on distinktiivinen ero, koska näiden avulla voidaan erottaa sanoja toisistaan (esim. [p**ii**] [p**uu**]). Edellisestä seuraa että suomessa katsotaan esiintyvän erilliset foneemit /i/ ja /u/. Sen sijaan esimerkiksi kahdella eri i-tyyppisen äänteen *varianteilla* [i] ja [I] ei eroteta merkityksiä suomessa. Täten sekä [i] että [I] kuuluvat samaan foneemikategoriaan /i/ – ne ovat foneemin /i/ *allofoneja*. Toisin on esimerkiksi englannissa, jossa on kaksi i-mäistä foneemia /i/ ja /I/ (ks. esim. Roach 1983, 15–18).

Puheen yhteydessä on paikallaan korostaa eroavuutta *foneettisen* ja *akustisen* eron välillä: Akustinen ero on kahden signaalin ero toisistaan. Kuulija ei välttämättä huomaa akustista eroa, mutta tietokone havaitsee tämän helposti. *Foneettinen* ero on äänteellinen ero, joka on riittävä, jotta äänteet luokitellaan kuuluviksi eri foneemiryhmiin, esimerkiksi äänteet [e] ja [i] tulkitaan foneemeiksi /e/ ja /i/. Jos kuulija pystyy havaitsemaan akustisen eron, mutta sillä ei ole kielellistä funktiota, puhutaan *foneemisesta* erosta. Äänteiden tunnistetusta erilaisuudesta ei kuitenkaan aina

seuraa, että kyseisen kieliryhmän kuulijat tulkitsisivat äänteitä erilaisiksi. Esimerkiksi suomalaiset tulkitsevat yleensä englannin [i] ja [I] -äänteet yhdeksi ja samaksi foneemiksi /i/ (Wiik 1965).

Siinä missä akustinen ero on puhtaasti signaalissa oleva ero, joka voidaan havaita koneellisesti täysin ongelmitta, foneettinen ero on sopimukseen perustuva abstrakti jako. Foneettinen ero perustuu sekä kuulijaan ääniperusteiseen tunnistamiseen että tietoon eri äänteiden tehtävästä tietyssä kielessä. Foneemiluokitus on siksi häilyvä. Esimerkiksi sama yksittäinen äänne voidaan äänneympäristöistä riippuen tulkita eri foneemeiksi, koska havaintoon vaikuttaa se, mitä henkilö olettaa kuulevansa. Kuulohavaintoon vaikuttaa paitsi äänneympäristö hyvin monet muutkin seikat. Esimerkiksi kuuluisa McGurk-koe osoittaa, miten näköhavainto suun liikkeistä vaikuttaa kuulohavaintoon (McGurk & MacDonald 1976).

Virhealttiutensa takia kieli onkin kehittynyt hyvin toisteiseksi. Toisteisuus tarkoittaa sitä, että kielessä käytetään paljon enemmän tilaa tietyn informaation välittämiseen kuin mitä käytetyn merkistön kannalta olisi välttämätöntä (ks. esim. Karlsson 1998, 62). Foneemitasolla tämä tarkoittaa esimerkiksi sitä, että missään maailman kielessä ei esiinny kaikkia foneemikombinaatioita distinktiivisesti, vaan vain osa eri ääntämystavoista esiintyy vapaassa variaatiossa keskenään (Ladefoged & Maddieson 1996). Sana- ja lausetasolla toisteisuutta edustaa muun muassa sanajärjestyksen merkitys sekä kongruenssi, eli taivutusmuotojen jakautuminen useammalle sanalle (”minun koira**ni**” tai ”kahdenne**lla**kymmene**llä**toise**lla** kerra**lla**”) (ks. esim. Karlsson 1998, 163–166).

Puheen koneelliseen käsittelyyn liittyy ennen kaikkea epätarkkojen kategorioiden käsittelyn ongelma. Tietokoneen avulla voidaan toki aina havaita akustisia eroja, mutta näiden luokittelu foneettisiksi eroiksi luotettavasti on äärimmäisen haastava tehtävä. Koneen kannalta ”kuulostaa samalle” ei siis ole ongelma, mutta päätös siitä, koska kaksi äännettä kuulostavat riittävässä määrin samalle, on sitäkin haastavampi.

3 Automaattinen puheentunnistus

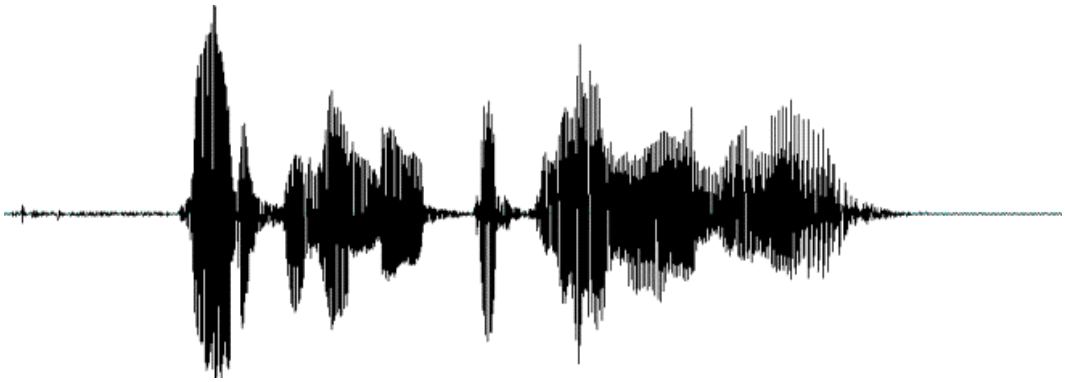
Automaattisessa puheentunnistuksessa tietokone analysoi puhesignaalia ja luokittelee puheesta erottuvat äänteet tietyn kielen foneemeiksi tai niistä koostuviksi puhutuiksi sanoiksi. Puheentunnistuksessa käytetyt menetelmät vaihtelevat, mutta yleisesti tunnistusprosessi on nelivaiheinen ja käsittää seuraavat tehtävät (vrt. Viikki 1999, 5–6):

- analoginen puhesignaali digitoidaan
- signaalista tunnistetaan osasävelten voimakkuudet
- signaalista tunnistetaan aikaikkunoittain äänteille tunnusomaisia piirteitä, ns. piirrevektoreita
- piirrevektoreiden avulla signaalin ääniä ryhmitellään tilastollisin menetelmin tiettyihin kategorioihin, esimerkiksi foneemeiksi.

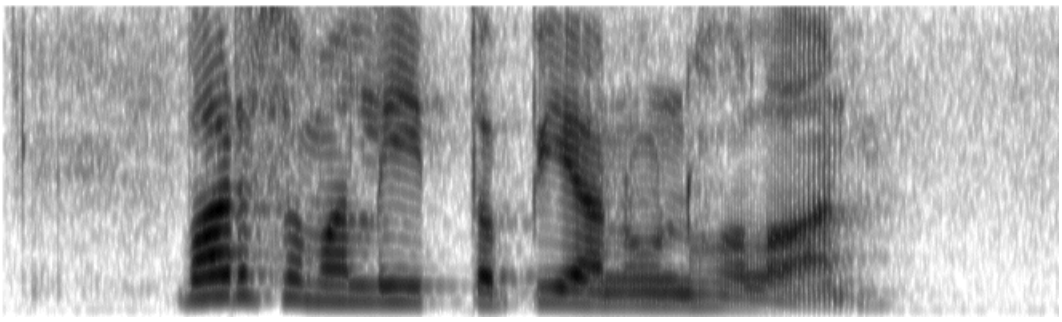
Puhesignaalin digitointia on käsitelty luvussa 2.3. Seuraavaksi käyn yleisluontoisesti läpi puheentunnistuksen muut vaiheet.

3.1 Puheentunnistusprosessi

Signaalin osasävelet ja niiden voimakkuudet saadaan selville, kun signaalille tehdään Fourier-analyysi. Spektri on esitys eri taajuuksien voimakkuuksista tietyllä ajan hetkellä (ns. frekvenssi-amplitudi-kuvaustapa). Spektrogrammilla voidaan yhdistää eri ajanhetkien spektrien informaatiota. Tällöin yksittäisten taajuuksien amplitudeja kuvataan yleensä joko tummuudella tai värillä. Kuviossa 1 on esitetty kompleksinen ääniaalto lauseesta ”Varsovan liitto pian unholaan” kirjoittajan puhumana. Kuviossa 2 puolestaan on esitetty puhesignaalin osasävelien taajuuksia kuvaava kapeakaistaspektrogrammi (taajuuksia kuvataan pystyakselilla ja aika kulkee vaaka-akselissa). Voimakkaimmat taajuudet erottuvat kuvassa tummempina.

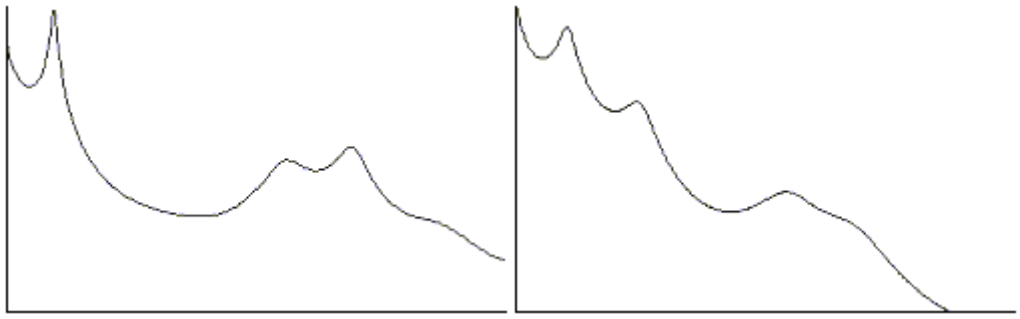


Kuvio1: Kompleksinen ääniaalto lauseesta: "Varsovan liitto pian unholaan."



Kuvio2: Spektrogrammi (kapeakaista) lauseesta: "Varsovan liitto pian unholaan"

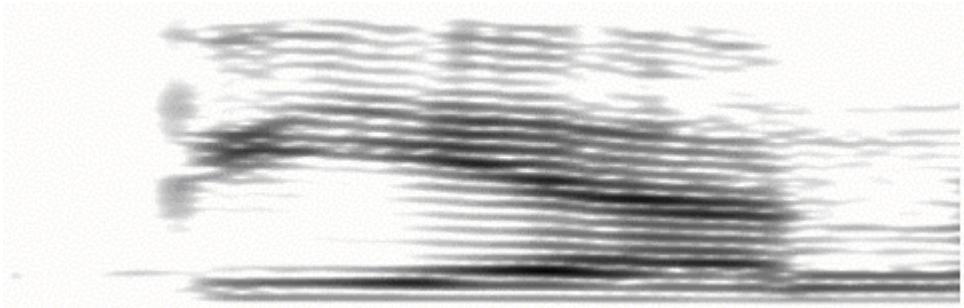
Spektristä voidaan tunnistaa erilaisia piirteitä, jotka ovat luonteenomaisia tietyille äänteille. Esimerkiksi tieto, onko äänne *soinnillinen*, eli värähtelevätkö äänihuulet sitä tuottaessa, on tärkeää informaatiota äänteen tunnistuksen kannalta. Soinnillisia äänteitä ovat kaikki vokaalit, sekä esimerkiksi konsonantit /l/, /m/ ja /g/. Soinnillisissa äänissä voidaan analysoida ääniväylän aiheuttamia muutoksia kurkunpää-ääneen.



Kuvio 3. Spektrit vokaaleista /i/ (vasemmalla) ja /u/ (oikealla).

Taajuushuippujen sijainnin perusteella voidaan päätellä, mikä äänne on kyseessä. Kuviossa 3 on esitetty (LPC-siloitetut) spektrit vokaaleista /i/ ja /u/.

Myös soinnittomilla äänillä on spektrissä tunnusomaisia piirteitä. Frikatiivit ovat konsonantteja, joissa ilman kulkua estetään siten, että syntyy suhiseva ääni (esim. /s/ /f/). Klusiilit puolestaan ovat konsonantteja, joissa ilman kulku hetkellisesti estetään kokonaan. Tästä seuraa aika-taajuus-spektrissä aukko, eli hetki hiljaisuutta. Pelkästä spektrin aukosta ei kuitenkaan voisi sanoa, onko kyseessä esimerkiksi /p/ vai /t/, vaan erotteluun tarvitaan myös tietoa suun liikkeestä juuri ennen sulkeumaa ja toisaalta heti sulkeuman jälkeen. Tämä ns. transitio, eli ääniväylän muodonmuutos asennosta toiseen, näkyy esimerkiksi kapeakaistaspektrissä taajuusviirujen nousuina tai laskuina. Aivan kuvion 4 alussa erottuu taajuusviirujen nousu (tumma alue), kun suun asento siirtyy äänteen /p/ edellyttämästä asennosta /i/:n vaatimaan asentoon.



Kuvio 4: Spektrogrammi (kapeakaista) lausahduksesta "pian".

Tunnusomaisista piirteistä voidaan muodostaa tietylle signaalin aikaikkunalle ns. piirrevektori (acoustic feature vector). Normaalisti signaalille muodostetaan piirrevektoreita 10ms välein (Robinson 1998). Signaalin piirrevektoreita hyödynnetään, kun puhesignaalin äänteitä analysoidaan ja tuotetaan tunnistustulos. Tässä vaiheessa signaalista havaittuja tunnusomaisia piirteitä verrataan puheentunnistimen kielimalliin, joka on eräänlainen säännöstö siitä, miten tietyt tunnusomaiset piirteet liittyvät tiettyihin äännekategorioihin. Vertaamalla tunnistettavien äänteiden piirteitä kielimalliin, ne voidaan luokitella kuuluviksi johonkin kielijärjestelmän sisältämään kategoriaan.

Tunnistus- ja luokitteluprosessi voidaan toteuttaa monella eri tavalla. Puheentunnistuksessa käytetään usein kätkettyjä Markovin malleja (Hidden Markov Model, HMM), jotka ovat todennäköisyyteen perustuvia tilakoneita. Puheen mahdollisista äänteistä muodostetaan tilakoneita, jossa tietty piirrevektori edustaa siirtymää tilasta toiseen. Siirtymien todennäköisyydet määrittyvät tunnistimen

koulutusvaiheen perusteella. Silloin tunnistimelle syötetään joukko lausahduksia sekä niiden oikeita tulkintoja, joiden perusteella tunnistin muokkaa tilastollisia mallejaan. Varsinaisessa tunnistuksessa tutkittavalle äänteelle lasketaan todennäköisyys, joka muodostetaan kun havaitut piirrevektorit liikkuvat tilakoneen läpi. (Ks. esim. Viikki 1999, 10–15.) Puheen piirrevektoreiden sovittamiseen foneemeiksi on käytetty muitakin menetelmiä, kuten neuroverkkoja (Robinson, Hochberg & Renals 1994).

Puhe on jatkuvaa signaalia, jonka vaihteluun vaikuttavat fyysiset suun ja nielun liikkeet. Tiettyä tarkasteluhetkeä edeltävät ja sitä seuraavat äänteet voivat auttaa tietyn äänteen tunnusomaisia piirteitä analysoitaessa. Tilakoneen avulla voidaan ottaa huomioon puheen jatkuva luonne ja tutkittavan piirrevektorin yhteensopivuus sekä sitä edeltäviin että seuraaviin piirrevektoreihin. Tilakoneita voidaan muodostaa yksi jokaista kielen foneemia kohden. Usein on kuitenkin tapana muodostaa *kontekstisidonnaisia* tilakoneita, joissa huomioidaan enemmän kuin yksi äänne kerrallaan. Puhetta mallinnetaan tällöin kahden (diphone) tai kolmen (triphone) peräkkäisen äänteen avulla, jolloin puheen jatkuva luonne voidaan paremmin ottaa huomioon. (Viikki 1999, 10–15.)

3.2 Puheentunnistuksen vaihtoehdot

Puheentunnistusjärjestelmät ovat kielisidonnaisia. Tämän takia englanninkielinen puheentunnistin ei sovellu suomenkielen tunnistamiseen, vaikka siihen vaihdettaisiinkin suomenkielinen sanasto. Tämä johtuu siitä, että eri kielten käyttämät foneemiaakkostot vaihtelevat. Kielet eroavat myös sen suhteen, mihin seikkoihin niiden käyttäjät kiinnittävät huomionsa arvioidessaan kahden eri äänteen samankaltaisuutta. Lisäksi yksittäisten äänteiden sallittu vaihteluväli voi vaihdella kielen mukaan, kuten myös äänteiden yleisyys puhutussa kielessä. (Ladefoged & Maddieson 1996.) Suurin osa nykyisin markkinoilla olevista puheentunnistimista on kehitetty tunnistamaan englanninkielistä puhetta. Puheentunnistimet voidaan jakaa eri kategorioihin sen perusteella ovatko ne:

- sanakirjapohjaisia vai ilman sanakirjaa toimivia
- jatkuvan vai diskreetin puheen tunnistimia
- puhujasta riippuvia vai riippumattomia

Sanakirjapohjaisessa tunnistimessa signaalin äänteet pyritään sovittamaan tunnistimen sanakirjassa oleviin sanoihin. Yleensä tunnistuksessa on välivaihe, jossa

äänteet ensin sovitetaan foneemikategorioihin ja vasta tämän jälkeen äänteet sovitetaan sanakirjan (*leksikon*) sanoihin. Sanakirjapohjaisen puheentunnistuksen tuloksena syntyy tunnistimen leksikon sanoista koostuvaa tekstiä. Tunnistuksen virheet ovat väärin tunnistettuja sanoja. Esimerkki virheestä on, että jokin puheessa esiintynyt sana on korvautunut väärällä leksikon sanalla. Tunnistin voi myös tulkita useamman puheessa esiintyneen peräkkäisen sanan yhdeksi sanaksi tai toisinpäin. Sanarajatkin voidaan tulkita väärin, kuten kahden erillisen sanan loppu- ja alkuosan tulkinta omaksi sanakseen.

Sanojen sovittamisessa sanakirjaan voidaan hyödyntää kielimalleja, koska sanat esiintyvät kielessä todennäköisesti tietyissä järjestyksissä ja jotkut järjestysvaihtoehdoista ovat kielellisesti mahdottomia. Esimerkiksi sanat ”I want to buy cheese” seuraavat toisiaan useammin edellä mainitussa järjestyksessä kuin ”want buy to I cheese”. Sanajärjestysten todennäköisyydet voidaan kouluttaa tunnistimelle ja parantaa sillä oikeintunnistamisen todennäköisyyttä.

Jos tunnistin **ei käytä sanakirjaa** apunaan, tunnistustulos muodostuu äännejaksoista, jotka on tuotettu piirrevektoreiden sovituksesta tunnistimen kategorioihin. Äännetunnistuksen perusta on, että tunnistettavassa kielessä on käytössä tietty (rajattu) määrä äänteitä, joiden avulla puheessa ilmaistaan kaikki kielen sanat. Tutkimusten raportoinnissa käytetään toisinaan sanaa *foneemitunnistin* (phoneme recognizer), vaikka usein kyse tarkasti ottaen on fooni- eli *äännetunnistimista* (phone recognizer). Esimerkiksi englannissa on arviolta vajaa 40 foneemia (tarkasta arvosta ei ole yksimielisyyttä, ks. esim. Roach 1983), mutta usein tunnistimet kuvaavat tunnistamaansa puhetta paljon suuremmalla joukolla erilaisia äänteitä. Äänteiden lukumäärää selittää se, että osa tunnistimista erottelee saman foneemin eri allofoneja tunnistusvaiheessa. Myös erikseen mallinnetut diftongit lisäävät tunnistettavien äänteiden määrää. Tunnistettavasta kielestä yleensä riippuu, käytetäänkö äänteiden luokitteluun foneemeja, allofoneja vai jotakin muuta jaottelua. Yhdenmukaisuuden vuoksi käytän tästä eteenpäin termiä äännetunnistin kattamaan nämä kaikki vaihtoehdot.

Äännetunnistimen virheet ovat väärin tulkittuja yksittäisiä äänteitä, eivät siis kokonaisia sanoja. Mahdollisia virheitä ovat korvaus, poisto ja lisäys. Vaikka sanakirjapohjaiset tunnistimetkin useimmin käyttävät äännetunnistusta ennen puheen lopullista sovittamista sanoiksi, sanoiksi sovittaminen auttaa yleensä korjaamaan joitakin äännetunnistuksen virheistä. Tämä johtuu siitä, että kielessä – saati sitten

tunnistimen leksikossa – ei ole olemassa loputtomasti sallittuja sanoja. Esimerkiksi jos tunnistin tuottaisi äännejakson [sulmemn], tämän sovittaminen sanakirjaan saattaisi hyvinkin tuottaa oikean tulkinnan ”suomen” vaikka äännetason tunnistuksessa on virheitä.

Puheentunnistin voi myös olla eräänlainen edellisten vaihtoehtojen välimuoto. Kuten edellä todettiin, jokaisessa kielessä on rajoitteita sille, miten eri äänteet voivat seurata toisiaan siten, että niistä muodostuu sallittu sana. Sen sijaan, että tunnistimen leksikko muodostuisi sanoista, se voidaan muodostaa kielessä käytössä olevista sanojen osista, kuten tavuista tai tavujen kombinaatioista (ks. esim. Whittaker, Thong & Moreno 2001). Tällaista osasanaleksikkoa käyttämällä voidaan sulkea pois joukko mahdottomia äännejärjestyksiä, esimerkiksi suomessa ”vf” tai ”frs” tai ”palä”. Tunnistimessa voidaan myös käyttää näiden sanoja pienempien osasanojen järjestykselle perustuvia kielimalleja (vrt. yllä).

Jatkuvan puheen tunnistin pystyy käsittelemään puhesignaalia, jossa on jatkuvaa puhetta. **Diskreetin puheen** tunnistin vaatii, että jokainen tunnistettava sana lausutaan erikseen, jolloin tunnistimen ei tarvitse käsitellä mahdollisesti useammasta sanasta koostuvia kokonaisuuksia kerrallaan. Jatkuvan puheen tunnistamisen ongelma on, että tunnistin ei kunakin hetkenä tiedä, minkä pituisesta äännejonosta tutkittava sana muodostuu. Tämän takia eri sanat voivat yhdistyä tai toisaalta yksi sana hajota useammaksi erilliseksi sanaksi (esimerkiksi ”Hello Kate” -> ”locate”).

Nauhoitetun jatkuvan puheen käsittelyyn tarvitaan jatkuvan puheen tunnistin. Diskreetin puheen tunnistinta tällaisen puheen käsittelyyn voidaan käyttää vain siinä tapauksessa, että sanat pystyttään automaattisesti eristämään signaalista ja tämän jälkeen syöttämään eteenpäin tunnistimelle. Puheessa ei kuitenkaan tavallisesti käytetä säännöllisesti taukoja sanojen välillä. Siksi sanojen eristäminen omiksi yksiköikseen vaatisi jo itsessään sanojen tunnistamista.

Puhujasta riippuva tunnistin muokkaa tunnistimen tilastollisia malleja tietylle henkilölle sopiviksi, eli tunnistin koulutetaan tunnistamaan tietyn henkilön puhetta. Tällöin puheentunnistimen käyttöönottoon liittyy se, että tunnistinta käyttävä henkilö puhuu ääneen tiettyjä sanoja tai lauseita, joiden oikea tunnistaminen on ilmaistu tunnistimelle etukäteen. Näiden ääninäytteiden perusteella puheentunnistimen tilastollista mallia tarkennetaan nimenomaan kyseisen henkilön puhetapaan soveltuviksi. **Puhujasta riippumaton** tunnistin soveltaa kaikille puhujille samaa

tilastollista mallia. Koska eri ihmisillä on erilaiset puheominaisuudet, kuten eri korkuinen ääni ja eri pituiset ääniväylät, tuottaa puhujasta riippuva tunnistin yleensä paljon vähemmän virheitä kuin puhujasta riippumaton (Viikki 1999, 39–40; Padmanabhan & Picheny 2001).

3.3 Kaupalliset tunnistimet

Useimmat kaupalliset räätälöimättömänä valmistuotteena myytävät jatkuvan puheen tunnistimet perustuvat sanakirjatunnistukseen. Tämä johtuu pitkälti niiden käyttötarkoituksista – kaupallisesti puheentunnistusta tarjotaan lähinnä sanelujärjestelmien ja toisaalta tietokoneiden puheohjauksen muodossa. Edellä mainituissa käyttötarkoituksissa ei niinkään ole hyötyä puheen äännejonon kuvailusta, vaan tunnistusprosessin lopullisena tavoitteena on muuntaa puhe tekstiksi tai tunnistaa annettu komento.

Sanakirjapohjaisessa tunnistuksessa signaalin akustista informaatiota käytetään hyväksi sovitettaessa signaalin äänteitä kokonaisiin sanoihin yksittäisten äännekategorioiden sijasta. Vaikka sanakirjapohjaiset tunnistimet tunnistavat yleensä ensin signaalista äänteitä ja vasta tämän jälkeen sovittavat näitä sanoiksi, pidempi analysoitava signaalisekvenssi sisältää enemmän informaatiota kuin yksittäinen aikaikkuna (vrt. Shannon 1948). Tämän takia sanakirjapohjainen tunnistus tuottaa yleensä luettavampaa tunnistustulosta kuin äännetunnistukseen perustuvat tunnistimet. Mikäli signaalissa puhutut sanat esiintyvät sanakirjassa, sanakirjapohjainen tunnistus myös tuottaa vähemmän virhetulkintoja kuin pelkkien äänteiden perusteella tehtävä tunnistus. Sen sijaan, jos puhuttu sana ei esiinny leksikossa, sanakirjapohjainen tunnistin tunnistaa sen väärin toiseksi leksikossa esiintyväksi sanaksi.

Sanakirjapohjaisten tunnistimien tunnistustarkkuuteen vaikuttaa sanakirjassa olevien sanojen määrä. Mitä enemmän siinä on sanoja, sen enemmän on myös vaihtoehtoisia tulkintoja samanmittaisille ja samankaltaiselta kuulostaville signaalisekvensseille. Tietokoneiden puheohjaukseen tarkoitetuissa tunnistimissa saattaa olla hyvin pieni sanavarasto, esimerkiksi 1000–2000 sanaa. Markkinoitavilla laajan sanavaraston tunnistimilla on yleensä valmis noin 40 000 sanan leksikko. Lisäksi usean tunnistimen sanastoa on mahdollista laajentaa noin 20 000 käyttäjän määrittelemällä sanalla. Laajan sanavaraston tunnistimien tunnistustarkkuus, eli oikein tunnistettujen sanojen prosentuaalinen määrä, liikkuu ihanteellisissa olosuhteissa 70–

90% välimaastossa. Pienen sanavaraston tunnistimissa tunnistustarkkuus saattaa olla lähes 100%. (Padmanabhan & Picheny 2001.)

Tunnistustarkkuuteen vaikuttaa luonnollisesti myös käyttäjän puhetapa. Jos esimerkiksi puhuja on hyvin nuori tai iäkäs tai puhuu murtaen, tunnistustarkkuus heikkenee huomattavasti. Myös taustahäly vaikuttaa negatiivisesti tunnistustarkkuuteen, koska tällöin signaaliin liittyy muita ääniaaltoja, joita on hankala erottaa puheen äänistä. (Padmanahan & Picheny 2001.) Lisäksi tunnistustarkkuuteen vaikuttaa signaalin laatu, siksi esimerkiksi puhelinlinjojen yli välitettyä ääntä on haastavampi käsitellä kuin korkeammilla näytteenottotaajuuksilla nauhoitettua ääntä (ks. luku 2.3).

3.4 Suomenkielinen puheentunnistus

Tällä hetkellä markkinoilla on kaksi suomenkielistä puheentunnistusjärjestelmää: Philipsin FreeSpeech-ohjelman suomenkielinen versio ja Lingsoftin markkinoima tunnistin. FreeSpeech-järjestelmän kehitys tosin on jo lopetettu, mutta järjestelmän suomenkielistä tunnistinta myydään vielä verkon kautta.[1] Philipsin FreeSpeech-tunnistimen toimintaperiaatteista tai tunnistustarkkuudesta ei ollut saatavilla tarkempia tietoja.

Lingsoftin tunnistin on puhujariippumaton foneemitunnistin, johon asiakas itse voi liittää vajaasta 2000 sanasta koostuvan sanaston. Järjestelmä soveltuu siten esimerkiksi tietokoneen ohjaukseen. 1800 sanan sanastolla Lingsoftin tunnistimen virheprosentit ovat 2-6% luokkaa (Lahti 2002).

Tietääkseni ei ole olemassa tutkimusta, jossa verrattaisiin eri kielisten puheentunnistusjärjestelmien laatua ja erityisesti tutkittaisiin sitä, onko olemassa joitakin kieliä, joiden tunnistus puheentunnistimen avulla on helpompaa kuin muiden kielien kohdalla. On kuitenkin usein ehdotettu, että sellaisia kieliä kuten esimerkiksi suomi ja italia, joissa puhuttu kieli lausutaan kutakuinkin samalla tavalla kuin se kirjoitetaan, olisi helpompi tunnistaa kuin esimerkiksi englantia. Puheentunnistuksen kannalta ei kuitenkaan ole tärkeää miten kielen kirjoitusasu liittyy sen lausuntaan. Sen sijaan puheentunnistuksen laatu vaihtelee sen mukaan, miten varmasti äänteiden erottelu puhesignaalista tapahtuu.

[1] Konttorityö-liike, katso http://www.konttorityo.fi/freespeechviva/ Saatavuus tarkistettu: 1.4.2003

Koska suomessa on vähemmän foneemeja, eli pienempi joukko vaihtoehtoisia tulkintoja, voisi olettaa suomen olevan helpommin tunnistettavissa kuin esimerkiksi englanti. Lisäksi suomenkieliset vokaalifoneemit ovat englannin vokaaleita järjestelmällisemmin samanlaisia äänneympäristöstä riippumatta (Wiik 1965, 145.)

Suomen ääntämisessä on kuitenkin sekä vokaalien että konsonanttien ääntämisessä käytössä pituusoppositio, jota englannissa ei (yksinään) esiinny (vrt. suomessa taka-, takaa, takka, takkaa, taakka ja taakkaa). Kaksoiskonsonanttien ja kaksoisvokaalien tunnistaminen voi puolestaan vaikuttaa puheentunnistuksen laatua heikentävästi. Pituusopposition tuottaminen perustuu nimittäin suhteelliseen äännepituuteen tietyssä äänneympäristössä, eikä absoluuttiseen aikaan (Wiik 1965, 112–113).

Puheentunnistuksen laatuun vaikuttaa myös sanojen erilaisuus sekä sanojen määrä. Suomen katsotaan kuuluvan ns. agglutinoivien kielien perheeseen, mikä tarkoittaa sitä, että suomessa aikaa, paikkaa ja muotoa ilmaistaan liittämällä päätteitä sanan loppuun (ks. luku 4.2). Näin esimerkiksi suomenkielisen sanakirjapohjaisen tunnistimen sanavarasto täyttyisi nopeasti saman sanan eri taivutusmuodoista, kun taas esimerkiksi englannissa sanojen taivutettuja muotoja on suhteellisen vähän. Gauvain, Lamel ja Adda (2000) vertasivat saksan ja ranskan leksikkojen kokoeroja englantiin verrattuna. He huomasivat, että 65 000 sanan sanakirja riitti kattamaan 99% heidän tutkimusaineistonaan käyttämänsä englanninkielisessä uutismateriaalissa käytetyistä sanoista, mutta osuudet olivat ranskalle 97,5% ja saksalle 95%. Syyksi he esittävät nimenomaan saksan ja ranskan kielten taivutusmuotojen, johdoksien ja yhdyssanojen runsauden (Gauvain, Lamel & Adda 2000). Nimenomaan suomen ja englannin leksikon kokoja vertailevia tutkimuksia ei tietääkseni ole tehty, mutta Alkula (2000) huomasi, että sanojen saattaminen perusmuotoon ennen käänteistiedostoon kirjoittamista pudotti käänteistiedoston kokoa alle puoleen alkuperäisestä (Alkula 2000, 152). Tämä tarkoittaa, että aineistossa noin puolet kaikista kirjoitusasultaan erilaisista sanoista olivat joidenkin toisten aineistossa esiintyvien sanojen taivutusmuotoja.

Puheentunnistuksessa voidaan huomioida taivutusmuotojen runsauden asettamat vaatimukset tunnistusprosessin luonteelle. Geutner ym. (1998) kehittivät puheentunnistusjärjestelmän serbo-kroatialle, joka suomen tavoin on vahvasti taipuva kieli. He sovelsivat tunnistusprosessin aikana mukautuvaa sanastoa. Tunnistus perustui kaksiosaiseen tunnistusprosessiin. Ensimmäisellä tunnistuskerralla käytettiin

tunnistimen perusleksikkoa. Tämän esitunnistustuloksen perusteella tunnistimelle valittiin uusi leksikko. Leksikko muodostettiin siten, että uuteen leksikkoon poimittiin sellaisten sanojen taivutusmuotoja, joista esitunnistustuloksessa esiintyi heikosti tunnistettuja variantteja. Nämä sanat vaihdettiin alkuperäisen leksikon vähemmän käytettyjen sanojen tilalle. Tämä jälkeen materiaali tunnistettiin uudelleen muokatun leksikon avulla. Järjestelmän avulla leksikon ulkopuolisten sanojen määrää väheni tekstistä riippuen 35–45% ja tunnistustulokset paranivat lähes 10%. (Geutner ym. 1998.)

4 Puhuttu luonnollinen kieli tiedonhaussa

Puhe on olennainen osa viestintäämme. Puheen avulla välittämme tietoa toisillemme siirtämällä käsitteitä äänen avulla kuulijan vastaanotettavaksi. Ihmiset käyttävät jokapäiväisessä viestinnässään *luonnollista kieltä*, erotuksena formaaleista kielistä (esimerkiksi matematiikka). Luonnollinen kieli on sosiaalisen vuorovaikutuksen tulos, joka muuttuu ja muovautuu käyttäjiensä mukaan. (ks. esim. Karlsson 1998, 1–5.)

Tiettyjen normien rajoissa ihmiset voivat käyttää kieltä hyvin eri tavoin ja silti tulla ymmärretyksi. Tästä seuraa, että eri ihmisillä yleensä on monipuolinen keinovalikoima puhua samoista asioista käyttämällä eri sanoja. Viestinnälliseen tyyliin jopa kuuluu, ettei samoja ilmaisukeinoja käytetä yhä uudestaan edes omassa puheessa.

Käsittelen tässä työssä sisältöperusteista puhetiedonhakua eli tiedonhakua, joka perustuu siihen, mitä haettavissa dokumenteissa on sanottu. En käsittele esimerkiksi hakujärjestelmiä, jotka hakevat dokumentteja muiden kuin sisältöperusteisten kriteerien perusteella. Muita kriteereitä voisivat olla puhujan henkilöllisyys, dokumenttien viemä muistitilan määrä tai dokumenttien luontiaika. Sisältöperusteisen tiedonhaun tavoite on tarjota käyttäjälle mahdollisuus löytää vaikkapa kaikki ne hakujärjestelmän piirissä olevat dokumentit, joissa on käsitelty suomalaista juustontuotantoa.

4.1 Tiedonhaku esiintymätasolla

Sisältöperusteisen tiedonhaun suurin ongelma on se, että dokumenteissa esiintyvää tietoa etsitään syntaksiin, eli esitystapaan perustuvien keinojen avulla. Tiedonhakuun ryhtyvä ihminen on kuitenkin kiinnostunut dokumenttien merkityssisällöstä. Tiedonhakujärjestelmän pitää siis kyetä tarjoamaan käyttäjälle sisällöltään kiinnostavia dokumentteja niissä esiintyvien merkkien perusteella. Järvelin (1993) havainnollistaa tätä problematiikkaa kolmen tason avulla: Hakijaa kiinnostaa dokumenttien (1) käsitetason sisältö, puhehaun tapauksessa ne aihepiirit joita puhedokumentit käsittelevät. Dokumenteissa nämä käsitteet on esitetty (2) ilmaisutasolla, puhedokumenteissa puhutun luonnollisen kielen avulla. Tiedonhakujärjestelmä käsittelee näitä ilmauksia (3) esiintymätasolla, eli vertaamalla käyttäjän hakupyynnöstä

saatujen hakuavaimien ja dokumenttien merkkijonoja (esimerkiksi puheentunnistimen äännejonoja) keskenään (mt.).

Luonnollisen kielen semanttisten rakenteiden automaattinen käsittely on vanha tietotekniikan kehittäjien haave. Luonnollisen kielen käsittely onkin edennyt valtavasti vuosikymmenten ajan. Silti ei voida sanoa, että tietokone ymmärtäisi, mitä kielellisesti ilmaistaan. Tiedonhaun perimmäinen ongelma on yksinkertaistettuna se, että tietokone käsittelee semanttista sisältöä syntaksiin perustuen. Tiedonhakujärjestelmiä voidaan kuitenkin kehittää ottamaan huomioon tiettyjä semanttisia piirteitä niiden ilmaisussa käytettävän syntaksin säännönmukaisuuksiin perustuen. Tämän avulla voidaan (ellei ratkaista niin ainakin) minimoida tiettyjä tiedonhaun ongelmia.

Yleisesti voidaan todeta, että sekä puhe- että tekstitiedonhaun kannalta ongelmallisia luonnollisen kielen piirteitä ovat monitulkintaisuus, synonyymien käyttö, sanojen morfologinen käyttäytyminen, ellipsien (poisjätettyjen elementtien) ja anaforien (esim. erilaisten pronominien) avaaminen (eli viittaussuhteiden ymmärtäminen) sekä kielen jatkuva kehittyminen. (Järvelin 1995, 165–166.) Näiden ongelmien ratkaisemiseksi on tehty paljon töitä erityisesti tekstitiedonhaun alueella. Tuloksena on syntynyt menetelmiä, joiden avulla luonnollisen kielen vaikutuksia voidaan ottaa huomioon hakujärjestelmissä. Ratkaisujen ei tarvitse olla aina monimutkaisia. Esimerkiksi yksi yleisimmin käytetyistä menetelmistä, hakukaavioiden muodostaminen, on todennäköisesti peräisin komentorivipohjaisista käyttöliittymistä. Näissä oli tärkeää pystyä käsittelemään isoja tiedostomääriä kerralla, esimerkiksi siirtämään joukko tiedostoja niiden nimen alkuosan perusteella. *Hakusanakaaviot* ovat olleet käytössä myös aikaisimpien täystäsmäyttävien (exact matching) hakujärjestelmien yhteydessä. Tällöin esimerkiksi on sallittu että hakusanojen loppuosa vaihtelee, jolloin yhdellä hakusanalla saadaan katettua sekä verbi että sen substantiivijohdannaiset (esim. **teach*** löytää myös **teach**er **teach**ing). Toinen esimerkki on jokerimerkin käyttö sanan keskellä. Tällä voidaan (joidenkin sanojen kohdalla) käsitellä esimerkiksi hakusanan vartalon taipumista (esim. **pöy?ä*** -> pöy**t**ä pöy**d**än).

Joissakin tiedonhakujärjestelmissä on myös käytetty ns. *kyselyn laajentamista* (Query Expansion ks. esim. Efthimiadis 1996). Tällöin hakujärjestelmän rinnalla olevasta sanakirjasta haetaan hakusanalle esimerkiksi synonyymejä, joita käytetään haussa alkuperäisten sanojen lisäksi. Menetelmän johdosta tiedonhakijan ei aina

tarvitse itse keksiä kaikkia niitä termejä, joita dokumenttien tuottajat ovat käyttäneet hänen kiinnostuksen kohteenaan olevasta aiheesta.

4.2 Suomenkielen erityispiirteet tekstitiedonhaun kannalta

Tiedonhakujärjestelmien kehitys on painottunut englanninkielisen aineiston tarpeiden täyttämiseen ja tämän takia suomenkielen erikoispiirteistä seuraa tiedon tallennuksessa ja haussa ongelmia. Useat luonnollisen kielen ongelmista ovat yleisiä kaikelle tekstitiedonhaulle. Kuitenkin suomi poikkeaa esimerkiksi englannista siinä suhteessa, että hakusanojen runsas morfologinen variaatio vaikeuttaa tiedonhakua.

Luonnolliset kielet voidaan morfosyntaktisesti eli taivutuskäyttäytymiseltään jakaa isoloiviin kieliin, agglutinoiviin kieliin ja fuusiokieliin. Isoloivissa kielissä (esimerkiksi vietnam) jokainen morfeemi[2] on oma sanansa. Agglutinoivissa kielissä puolestaan morfeemit liitetään muuttumattomaan sanavartaloon. Tällainen kieli on esimerkiksi turkki. Fuusiokielissä morfeemien rajat eivät ole selvät, koska sanavartalo muuttuu kun siihen liittyy morfeemi. (Pirkola 2001.) Yleensä kielet eivät ole puhtaasti minkään tyypin edustajia: esimerkiksi suomi voidaan katsoa kuuluvan agglutinoiviin kieliin, vaikka siinä päätettä edeltävässä sanavartalossa voi tapahtua muutoksia. Englanti on puolestaan isoloivampi kieli kuin suomi, vaikka siinäkin esimerkiksi monikkoa ilmaiseva morfeemi liitetään suoraan sanavartaloon (one dog, three dogs). (Itkonen 2001, 64–94.)

Toisin kuin englanninkielessä, esimerkiksi aikaa, paikkaa ja syytä ilmaisevat määreet liitetään suomessa suoraan sanavartalon loppuun (esimerkiksi talo, talossa, taloon, vrt. a house, in the house, into the house). Sanavartalo voi kuitenkin vaihdella ja näiden muutosten takia seuraa tiedonhaussa ongelmia. Usein tiedonhakujärjestelmissä on mahdollista katkaista hakusana, jolloin sallitaan, että hakusanan loppu vaihtelee. Suomessa sanan vartalo voi kuitenkin muuttua eri päätteiden mukaan (esimerkiksi työ, töiden). Tällöin sanoilla ei aina ole yhtä vartaloa, johon päätteitä liitetään.

Tekstitiedonhaun kannalta eräs tyypillisistä ongelmista on *homografia*, toisin sanoen kaksi merkitykseltään eri sanaa kirjoitetaan samalla tavalla. Tätä esiintyy myös englannissa (kuusi=puu, kuusi=numero vrt. stick=tarrautua, stick=puikko).

Suomenkielen taivutusmuodoista seuraa kuitenkin vielä lisäksi se, että kahden merkitykseltään eri sanan taivutusmuodot saatetaan kirjoittaa samalla tavalla (hauissa=kaloissa, hauissa=hakujen sisällä). Tällöin on kyse *taivutusmuotohomografiasta* (ks. esim Järvelin 1995, 168).

Suomessa on mahdollista johdoksien avulla tuottaa täysin uusia sanoja. Tällöin etu- tai jälkiliitteillä lisäämällä muutetaan kantasanan merkitys. Johtimen kantasana voidaan tunnistaa automaattisesti, mutta vaikeuksia aiheuttaa se, että johto-opissa on huomattavasti enemmän epäsäännöllisyyttä kuin taivutuksessa. Esimerkiksi johdin ele johtaa kantaverbistä uuden verbin, joka yleensä ilmaisee sen, että kantasanan teko toistetaan useita kertoja: 'hyppää' -> 'hyppelee'. Kuitenkin 'sanella'-verbillä on oma merkityksensä, joka on 'sanoa'-verbistä erillinen. Tässä tapauksessa johtimen semanttinen vaikutus kantasanaan ei päde. Sama pätee sanoihin 'tappaa' ja 'tappelee'. Lisäksi kaikilla johdoksilla ei välttämättä edes ole kantasanaa, näin esimerkiksi sanalla 'kirvellä'.

Suomessa on paljon yhdyssanoja ja niitä syntyy lisää siten, että vakiintuneita sanaliittoja aletaan kirjoittamaan yhteen. Yhdyssanoissa loppuosa ilmoittaa yleensä pääluokan ja alkuosa alaluokan. Tiedonhaun kannalta loppuosan tunnistaminen on siksi usein tärkeää ('rivi**talo**' tai 'maaenergia**talo**'). Englannin kielessä yhdyssanat kirjoitetaan useimmin sanaliittoina, toisinaan yhdysviivalla erotettuina, jolloin rajan vetäminen on helppoa (tai tarpeetonta) (Alkula & Honkela 1992, 19). Yhdyssanojen ongelmallisuus liittyy toisaalta myös siihen, että niiden merkitys ei välttämättä ole johdettavissa niiden osien merkityksestä (esim. 'vanha' ja 'poika' vs. 'vanhapoika'). Lisäksi myös homografiaa voi esiintyä yhdistämättömän sanan taivutusmuodon ja yhdyssanan välillä (esim. 'runoilta' vs. 'runo' ja 'ilta').

4.3 Suomenkielinen puhe tiedonhaussa

Puhehaun tutkimus on niin tuore tutkimusala, että siinä on toistaiseksi sovellettu lähinnä tekstitiedonhaun oppeja. Puhehaun pääasiallinen ongelma on ollut se, miten puhesignaalia voidaan saattaa sellaiseen muotoon, että siihen voi soveltaa tekstitiedonhaun menetelmiä. Menetelmät ovatkin lähinnä keskittyneet

[2] Perinteisesti morfeemi määritellään sanan pienimmäksi yksiköksi, jolla on oma merkitys. (ks. esim. Karlsson 1998, 94–95.)

virhesietoisten, eli vääriä tulkintoja sisältävän materiaalin käsittelyyn soveltuvien, tiedonhakumenetelmien kehittämiseen (ks. luku 5.2.2).

Tekstihaussa voidaan morfologisten tulkintaohjelmien avulla parantaa hakutulosten laatua täystäsmäytykseen perustuvissa järjestelmissä. Morfologiset tulkintaohjelmat tunnistavat sanan taivutusmuodot ja niillä pystytään esimerkiksi saattamaan tekstin taivutusmuotoiset sanat perusmuotoonsa. Haussa löydetään enemmän tärkeitä dokumentteja, koska hakusanan eri taivutusmuotojen esiintyminen tekstissä pystytään ottamaan huomioon. Haun tarkkuus eli haluttujen dokumenttien osuus kaikista hakujärjestelmän palauttamista dokumenteista kasvaa. Tämä johtuu siitä, että hakijan itse katkaisemat sanat tuottavat myös sellaisia sanoja, jotka eivät ole katkaistun sanan taivutusmuotoja, vaan esimerkiksi johdoksia. (Alkula 2000.)

Myös puhemateriaalin haun yhteydessä voidaan hyödyntää morfologisia tulkintaohjelmia, jos puhe ensin tunnistetaan tekstiksi. Jos puheentunnistamisessa käytetään sanastopohjaista lähestymistapaa lienee suurin vaikutus hakuun sillä, miten hyvin tunnistimen sanakirja kattaa puhedokumentissa käytetyt sanat. Kuten luvussa 3.3 totesin, suomenkielinen morfologia saattaa vaikeuttaa puheen sanakirjapohjaista tunnistusta, koska sanojen lukuisat taivutusmuodot kasvattavat sanakirjan kokoa. Toisaalta tiedonhaun kannalta on usein mielenkiintoista selvittää nimenomaan sanan kantasana, jolloin taivutusmuodon tunnistaminen oikein ei ole välttämätöntä. Taivutusmuodot saattavat kuitenkin vaikeuttaa myös oikean kantasanan tunnistamista. Näin voi käydä esimerkiksi niiden sanojen kohdalla, joissa sanavartalo muuttuu päätteen mukaan.

Tutkimuksessani käytetyssä materiaalissa (ks. luku 6.3), eli puhutuissa uutisissa on paljon sellaisia sanoja, joita tavallisessa sanakirjassa ei ole, kuten henkilöiden ja paikkojen erisnimiä. Jos puhetta tunnistetaan sanakirjapohjaisella tunnistimella, monet näistä tiedonhaun kannalta olennaiset sanat jäävät tunnistustuloksesta puuttumaan. Vaihtoehtona sanakirjapohjaiselle tunnistukselle puhemateriaali voidaan tunnistaa äännetunnistuksen avulla. Tämän jälkeen käytettävät osittaistäsmäytysmenetelmät sallivat paitsi väärintunnistettujen myös taivutusmuotoisten sanojen löytymisen dokumenteista.

Puhe eroaa tekstistä merkittävästi siinä, että puheessa esiintyvät sanat eivät ole samalla tavalla valmiiksi erotettu toisistaan. Sanarajojen tunnistamisen ongelma saattaa esimerkiksi johtaa siihen, että dokumentissa tulkitaan esiintyvän sanaraja sellaisessa kohdassa, jossa sitä alkuperäisessä puheessa ei ollut. Sanarajojen

tunnistaminen puheesta edellyttää nimittäin sanojen tunnistamista; vasta sanojen tunnistamisen jälkeen tunnistin voi päätellä niiden välissä olevan sanaraja.

Sanarajojen tunnistusvaikeuksista johtuen sanakirjapohjainen puheentunnistin saattaa myös tunnistaa puhemateriaalissa esiintyvät yhdyssanat joko yhdyssanoiksi tai sanaliitoiksi. Koska tiedonhaussa yleensä ollaan kiinnostuneita yhdyssanojen osista, sanojen tunnistaminen erillisiksi tuskin tuottaa erityisiä ongelmia. Ongelmallisiksi voivat osoittautua lähinnä sellaiset yhdyssanat, joista vain toinen osa pystytään tunnistamaan. Erityisen hankalaksi tämä voi muodostua, jos yhdyssanan merkitys ei ole johdettavissa sen osien merkityksestä.

Puheeseen perustuvaa tiedonhakua helpottaa se, että homografiset sanat eivät välttämättä puhuttuna kuulosta samalle. Tämä pätee varsinkin taivutusmuotohomografiaan. Vaikka kahden sanan (hauissa, haku; hauissa, hauki) kirjoitusasu onkin homografinen, kuulostavat nämä yleensä lausuttuina erilaiselle ([ha'uissa] [haui'ssa]). Tällöin puhuttu muoto voi auttaa merkityksen päättelyssä.

Puheessa homografian vastine on *homofonia*, joka tarkoittaa samalta kuulostamista. Homofoniassa kaksi merkitykseltään erilaisella sanalla on sama (oikeaoppinen) lausunta. Sekä suomen- että englanninkielessä homofonia liittyy usein homografiaan, jolloin se ei vaikuta puheentunnistukseen (stick, stick; kuusi, kuusi). Jonkinasteista homofoniaa saattaa kuitenkin esiintyä myös kahden kirjoitusasultaan erilaisen sanan välillä (esimerkiksi picture, pitcher; Virta, Wirta).

Suurin ongelma puheen käsittelyssä on kuitenkin se, että tietokone käsittelee puhetta signaalin akustisten erojen perusteella. Ongelma syntyy siksi, että sama sana saatetaan lausua hyvinkin eri tavoin muuttamatta sen semanttista sisältöä. Samalla kuitenkin myös kaksi merkitykseltään eri sanaa saatetaan lausua hyvin samankaltaisella tavalla. Koska jokainen puhunnos on erilainen, ”samalle kuulostaminen” ei juurikaan tuota koneelle ongelmia. Koneellisesti voidaan nimittäin havaita äänisignaalissa sellaisiakin pieniä muutoksia, joita ihmiskorva ei kykene erottamaan. Sen sijaan samankaltaisuuden määrittäminen enemmän tai vähemmän toisistaan eroavien äänien välillä on puheentunnistuksen kannalta haasteellinen tehtävä. Puheentunnistukseen liittyvät seikat on käsitelty luvussa 3.1 ja sekä suomenkielen erityisominaisuuksia puheentunnistuksen kannalta luvussa 3.3.

5 Puhetiedonhaun menetelmät

Puhehaulla tarkoitan tässä työssä sellaista tiedonhakua, joka kohdistuu puhumalla tuotettuun ja audiotallenneteiseen tietoon. Muunlaisetkin määritelmät ovat mahdollisia: puhehauksi voisi esimerkiksi nimittää kaikkea sellaista tiedonhakua, jossa hakupyyntö esitetään puheen avulla. Määrittelemällä puhehaku tarkoittamaan sellaista tilannetta, jossa haettava tieto esiintyy (ainakin alkuperäisessä muodossaan) puhutussa muodossa, sallitaan varsinaisen tiedonhaun tapahtuvan monilla eri tavoilla. Valitsemani määritelmä painottaa sitä, miten haettava tieto ja sen tallennusmuoto vaikuttavat tiedonhaun onnistumiseen. Tiedonhaun kannalta on luonnollisesti oleellista myös tiedonhakijan tukeminen hänen esittäessään tiedon tarpeitaan. Kuitenkaan tämä näkökulma ei ole käsillä olevan tutkimuksen kannalta oleellinen.

Puhemateriaalin hakuun vaikuttaa ensisijaisesti kaksi seikkaa: puhesignaalin jatkuva luonne sekä se, että puhe on luonnollista kieltä. Edellisissä luvuissa on käsitelty puhesignaalia ja signaalinkäsittelyyn liittyviä seikkoja sekä luonnolliseen kieleen kohdistuvan tiedonhaun ongelmia. Seuraavaksi esittelen, millaisia erilaisia lähestymistapoja puhehaussa on käytetty.

5.1 Viitetietokannoista automaattiseen sisältöperusteiseen tiedonhakuun

Audiotallenteita, muun muassa puhemateriaalia, on perinteisesti haettu metadatan perusteella. Metadata on tietoa tiedosta, eli puhetallenteen tapauksessa se on yleensä ollut esimerkiksi ihmisen tuottama transkriptio[3] tai kuvaus, joka on ilmaissut puhetallenteen sisällön. Yksi lähestymistapa puhedokumenttien tiedonhakuun on ollut kuvata tiedostoja sanallisesti joko asiasanoilla tai tiivistelmillä ja tämän jälkeen soveltaa tällä tavalla luotuihin tiedostoihin perinteistä tekstitiedonhakua. Tällöin audiomuotoisen tiedoston sisältöä kuvaa tekstimuotoinen esitys ja tiedonhaussa hakusanoja pyritään täsmäyttämään varsinaisten dokumenttien sijasta näihin tekstimuotoisiin kuvailutiedostoihin. Koska tiedonhaun menetelmät yllä kuvaillussa tapauksessa kohdistuvat tekstiin, ongelmaksi muodostuu ensisijaisesti se, miten

kattavia kuvailutiedostot ovat ja miten hyvin ne kuvaavat puhemateriaalia. Varsinaiseen täsmäyttämiseen voidaan soveltaa perinteisiä tekstitiedonhaun menetelmiä.

Kuvailutiedostoja hyödyntävällä lähestymistavalla on kuitenkin monia ongelmia: Ensinnäkin audiomuotoisen materiaalin tuotanto on nykyisellään niin nopeaa, etteivät henkilöresurssit aina riitä tuottamaan kuvailutiedostoja. Toiseksi tiedoston kuvailu siten, että kaikki tiedoston hakujen kannalta oleelliset sanat ja ilmaukset pystytään mainitsemaan kuvailutiedostossa, on äärimmäisen hankala – usein jopa mahdotonta. Kuvailutiedostojen luomisvaiheessa kun on käytännössä mahdoton ennakoida, mitkä seikat dokumentissa myöhemmin hakijoita kiinnostavat. Saattaahan esimerkiksi olla, että hakija ei ole lainkaan kiinnostunut dokumentin asiasisällöstä, vaan esimerkiksi puhujan käyttämistä vertauskuvista asian ilmaisussa.

Automaattisella tiedon tallennuksella ja haulla tarkoitetaan sitä, että haettavan informaation tallennukseen ja hakuun liittyvä tietojenkäsittely suoritetaan täysin automaattisesti, ilman inhimillistä kontrollia. Karkeasti ottaen tällöin on olemassa seuraavat kaksi eri lähestymistapaa puhehaun toteuttamiselle. Toinen perustuu sanakirjapohjaisen puheentunnistuksen, jonka avulla puhe tunnisetaan tekstiksi. Tunnistamisen jälkeen syntyneeseen tekstiin sovelletaan tekstitiedonhaun menetelmiä. Toisessa lähestymistavassa ei pyritä tuottamaan puheesta tekstiä, vaan puhetta käsitellään jollakin abstraktiotasolla, esimerkiksi puheen äänneasua foneemeina. Haussa hakusanat tuotetaan tähän samaan muotoon, jonka jälkeen hakuohjelmalla etsitään samankaltaisuuksia hakusanan ja tietokannan puhemateriaalin esitysmuotojen väliltä.

5.2 Sanakirjapohjaiseen puheentunnistukseen perustuva puhehaku

Luonnollinen lähestymistapa puhehakuun on soveltaa puhetallenteisiin sanakirjapohjaisia puheentunnistusmenetelmiä (ks. luku 3.1) ja tämän jälkeen käyttää pitkään kehitettyjä tekstitiedonhaun menetelmiä tunnistimen tuottamaan transkriptioon.

---

[3] Transkriptiolla tarkoitetaan tässä työssä jollakin ennalta sovitulla tarkkuudella tuotettua tekstimuotoista esitystä puhemateriaalista.

Sanakirjapohjainen tunnistus nousi vaihtoehdoksi, kun puheentunnistusmenetelmät alkoivat kehittyä 1990-luvun puolivälissä. Ensimmäisiä sanakirjapohjaiseen tunnistukseen liittyviä tutkimuksia edustaa vuonna 1995 Cambridgen yliopistossa tehty Video Mail Retrieval -projektin ensimmäisen vaiheen tutkimus. Projektissa käytettiin hyvin pienen sanavaraston puhujakohtaista puheentunnistinta yhdessä tekstitiedonhaun kanssa. Puheentunnistusjärjestelmä tunnisti videopostiviesteistä 35 avainsanaa, joiden perusteella viestejä haettiin pienestä 300 viestin tietokannasta (Spärck Jones ym. 1996).

Sanakirjapohjaiseen tunnistukseen perustuvaa (englanninkielistä) puhetiedonhaun tutkimusta on vuosina 1997–2000 tutkittu ja raportoitu ensisijaisesti TREC-konferenssin[4] yhteydessä. Vuonna 1997 nimittäin käynnistyi puhemateriaalin hakuun keskittyvä erityishaara (SDR-Track). Tutkimus tuotti niin hyviä tuloksia, että vuonna 2000 puhehaku (englanninkielinen) arvioitiin ratkaistuksi ongelmaksi ja katsottiin, että TREC-voimavarat jatkossa kannattaa keskittää haastavampien tehtävien ratkaisemiseen, kuten multimediatiedonhakuun (Garofolo ym. 2000).

Kautena 1997–2000 puheentunnistuksen laatu myös kehittyi huimasti. Ensimmäisen TREC-SDR -erityishaaran tutkijoille tarjotun ja valmiiksi tunnistetun testimateriaalin sanavirhemäärä (Story Word Error Rate tästä eteenpäin *SWER*) oli peräti 50%. Vuotta myöhemmin (1998) käytössä oli Carnegie Mellonilla kehitetty SPHINX-III -tunnistin, jonka kahden tunnistusversion sanavirhemäärät olivat 33,8% ja 46,6%. Viimeisenä TREC-SDR-vuotena kilpailleiden tunnistimien sanavirhemäärät olivat kaikki alle 25%. (Garfolo ym. 2000.)

5.2.1 TREC puhemateriaali ja sen tunnistaminen

Tiedonhakuun soveltuvan tunnistimen pitää pystyä tunnistamaan jatkuvaa puhetta. Lisäksi tiedonhaun tarkoituksiin ei ole järkeä soveltaa hyvin pienen sanaston tunnistinta, koska sanaston koko rajoittaa haettavien sanojen määrää.

[4] TREC (Text Retrieval Conference) on National Institute of Standards and Technologyn (NIST) jatkuva hanke, jossa rakennetaan suurta, kansainvälistä, yleisessä käytössä olevaa tiedonhaun tutkimusympäristöä. Tutkimusten yhteismitallisuuteen pyritään sillä, että jokainen tutkimuslaitos kohdistaa menetelmänsä yhteisiin testikokoelmiin ja kaikki käyttävät samoja hakupyyntöjä. Yhteisiä aineistoja käyttämällä pyritään helpottamaan eri tiedonhakumenetelmien välistä vertailua. Samalla saadaan myös arvokasta dataa esimerkiksi aineiston vaikutuksesta hakumenetelmien toimintaan. Katso esim. http://trec.nist.gov/ Saatavuus tarkistettu: 1.4.2003

Markkinoiden parhaimmat jatkuvan puheen puheentunnistimet perustuvat kouluttamiseen, jossa puheentunnistin on ennen käyttöönottoa koulutettu tunnistamaan sanoja erityisesti juuri tietyn puhujan puhetyylillä. Lisäksi tunnistimet yleensä edellyttävät, että puhe tapahtuu hyvin suotuisissa olosuhteissa, esimerkiksi taustahälyttömässä ympäristössä suoraan mikrofoniin. Puheen taustalla olevat muut äänet heikentävät tunnistuksen laatua huomattavasti.

Puhehaun tutkimuksissa materiaalina on usein käytetty vähemmän suotuisissa olosuhteissa tuotettua puhetta, esimerkiksi video- ja puhepostia (Spärck Jones ym. 1996) tai tv- tai TREC-konferenssin puitteissa radio-ohjelmavirtaa (Garofolo ym. 2000).

TREC-konferenssin puhehaun erityishaaran materiaalina käytettiin audiotallenteista uutismateriaalia. TREC-SDR -erityishaaran käynnistyessä vuonna 1997 valittiin arvioinnin helpottamiseksi tehtäväksi ns. known item search. Siinä tiedonhakutehtävä oli löytää yhteensä 50 etukäteen tunnettua uutista aineiston joukosta. Materiaalina käytettiin tiedonhaun näkökulmasta pientä tietokantaa, joka koostui noin 50 tunnista uutispuhetta.

Vuonna 1998 TREC:n dokumenttitietokanta kasvoi 87 tuntiin ja tehtävissä siirryttiin tunnetun uutisen löytämisestä perinteiseen tiedonhakuun. Siinä tavoitteena oli sekä löytää tietokannasta tiettyä aihepiiriä käsittelevät uutiset että listata ne saantilistan kärkeen (ranked retrieval). Vuosina 1999–2000 materiaalina käytettiin tiedonhakumielessä realistisempaa tietokantaa (550 tuntia puheuutisia), jossa varsinaisen puheen lisäksi oli myös mainoskatkoja ja taustamusiikkia. Vuosien 1999 ja 2000 varsinainen ero oli siinä, että jälkimmäisenä vuotena dokumenttien esikäsittelytehtäviä (esimerkiksi dokumenttien rajojen selvittäminen ja mainosten suodattaminen materiaalista) siirrettiin tunnistusjärjestelmän suoritettavaksi. Vuonna 1999 esikäsittelyn automatisointi oli vielä ollut tutkijoille vapaaehtoista. (Ks. esim. Garofolo ym. 2000.)

Yksi esimerkki TREC-tiedonhaussa käytetystä tunnistimesta on LIMSI:n[5] OLIVE-projektin yhteydessä kehitetty puheentunnistin, joka soveltuu erityisesti uutismateriaalin tunnistamiseen. Tunnistimella on kaupallisiin järjestelmiin verrattuna suhteellisen suuri (65 000 sanan) sanavarasto. Järjestelmällä päästiin melko hyvään tunnistustulokseen (SWER 21,5%) uutismateriaalin vaihtelevasta sisällöstä huolimatta. Hyviin tuloksiin päästiin siten, että puhemateriaali ensin eroteltiin

[5] Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur

akustisesti samanlaisiksi yksiköiksi, jonka jälkeen jokaisen yksikön tunnistamiseen käytettiin kyseisen tyyppisen materiaalin tunnistamiseen erikoistunutta tunnistinkomponenttia. Esimerkiksi mies- ja naispuhujille, puhelinkeskusteluille ja taustamusiikin päälle puhutulle puheelle oli kehitetty omat tunnistinkomponentit. (Gauvain, Lamel & Adda 2000.)

5.2.2 Tunnistusvirheet ja niiden käsittely

Ehkä suurin yllätys puhehaun tutkijayhteisölle on ollut se, että tunnistimien suuretkaan sanavirhemäärät eivät merkittävästi ole huonontaneet hakutulosta tekstihakuun verrattuna. Alkuoletuksena voisi kuvitella, että yksikin väärintulkittu sana voisi kaataa koko hakutuloksen. Kuitenkin on tavallista, että puhedokumentti, joka käsittelee jotakin tiettyyn sanaan liittyvää aihepiiriä, sisältää myös monta tämän sanan esiintymää. Täten esimerkiksi budjettia käsittelevä puheessa tuskin tyydytään mainitsemaan sana 'vero' vain kerran. Tässä tapauksessa ei haittaa, vaikka se puhedokumentin jossakin kohdassa olisi tunnistettu 'ero'-sanaksi. 50% sanavirhemäärä tarkoittaa sitä, että dokumentissa esiintyy myös 50% oikein tunnistettuja sanoja. Toisin sanoen näiden joukosta voi löytyä ainakin joku annetuista hakusanoista. Samoin puheessa käsitellään todennäköisesti muita aiheeseen liittyviä sanoja, vaikka tiettyä hakusanaa ei tunnistettaisikaan oikein. Esimerkiksi limnologiaa käsittelevä uutinen sisältää todennäköisesti myös sellaisia sanoja kuten 'biologia', 'meribiologia' ja 'kala'.

Allan (2002) esittää, että mitä pidempi dokumentti on, sitä todennäköisemmin se sisältää runsaasti toistoa ja rinnakkaistermejä, joista ainakin osa tunnistetaan oikein. Tämä selittäisi sen, miksi korkeatkaan sanavirhemäärät eivät vaikuta hakutuloksiin dramaattisesti. Heikko tunnistustulos saattaa kuitenkin tuottaa ongelmia erityisesti silloin, kun haettavissa dokumenteissa ei ole kovinkaan paljon toistoa. Tällöin on olemassa suuri todennäköisyys, että vain muutaman kerran dokumentissa esiintynyt hakusana on tunnistettu väärin. Sen tähden tunnistusvirheet voivat olla erityisen ongelmallisia lyhyiden dokumenttien, kuten puhepostiviestien haussa (Allan 2002).

Tunnistusvirheiden suurta vaikutusta nimenomaan lyhyiden dokumenttien haun laatuun indikoi myöskin tutkimus, jossa verrattiin puhuttujen hakukysymysten sanavirhemäärien vaikutusta haun onnistumiseen. Pidemmissä (noin 40-60 sanan kysymyksissä) korkeankaan (50%) sanavirhemäärän heikentävä vaikutus hakutulokseen ei ollut kuin 15% (verrattuna alkuperäisiin kyselyihin). Lyhyemmillä

kysymyksillä (5-8 sanaa) taas sama 50% sanavirhemäärä aiheutti 60% huonompia hakutuloksia, koska lyhyempi kysymys sisälsi vähemmän rinnakkaistermejä ja toistoa. (Barnett ym. 1997.)

Puheentunnistusjärjestelmien tuottamien melkoisten virhemäärien takia useat puhehaun tutkimuksista ovat keskittyneet väärintunnistettujen sanojen ongelman ratkaisemiseen. Seuraavaksi käsittelen pääasiassa TREC-SDR -erityishaaran toimintaan osallistuneiden tutkimuslaitosten raportoimia hankkeita ratkaista väärintunnistettujen sanojen aiheuttamia ongelmia.

Tunnistimen tuottamat todennäköisyysarvot

Jotkut puheentunnistimet tuottavat transkriptioiden lisäksi tunnistamilleen sanoille todennäköisyysarvoja. Ne kuvaavat sitä kuinka varma tunnistin on tunnistustuloksestaan (confidence value). Eräs lähestymistapa virheiden käsittelyyn on ollut hyödyntää näitä arvoja haun yhteydessä. Tällöin tutkittaessa sanan esiintymistä dokumentissa tarkastellaan myös siihen liitettyä ”oikean arvauksen” todennäköisyyttä. Vähemmän todennäköiset sanat katsotaan vähemmän painaviksi myös hakutulosta laskettaessa.

Tunnistimen antamien todennäköisyysarvojen hyödyntämisellä saavutetut hakutulokset eivät kuitenkaan ole olleet hyviä. Sanderson ja Crestani (1998) totesivat, että todennäköisyysarvojen hyödyntäminen itse asiassa laski hakutuloksen laatua. Tämä johtui siitä, että sanan todennäköisyysarvo liittyi enemmän sanan pituuteen kuin siihen, oliko se tunnistettu oikein vai väärin. Siegler (1999) kokeili todennäköisyysarvon normalisointia sanan pituuden suhteen, mutta tulokset eivät parantuneet.

Tunnistimien vaihtoehtoiset arvaukset

Vaikka tunnistin saattaa tarjota ensisijaiseksi tunnistustulokseksi väärän sanan, oikean tunnistustuloksen todennäköisyys saattaa poiketa vain vähän väärän tulkinnan todennäköisyydestä. Jos puheentunnistimelta saadaan yhden ainoan tunnistustuloksen sijasta lista tunnistimen parhaimmista arvauksista, voidaan näitä vaihtoehtoisia arvauksia myös hyödyntää tiedonhaussa. Siegler (mt.) kokeili seuraavaksi useamman tunnistushypoteesin käyttämissä pelkän ykköstulkinnan sijasta (n-best-lista, eli n parasta tunnistustulosta). Siegler käytti n-best-listoja, jotka

koostuivat noin 10 sekunnin mittaisten puhepätkien tunnistustuloksista. Pitkä aikaikkuna johti siihen, että erilaisiksi katsottiin jo sellaiset lauseet, jotka erosivat toisistaan vain yhden sanan tai tauon verran. Tulokset kuitenkin osoittivat, että ylimääräisten tunnistusvaihtoehtojen käsittely oli kannattavaa. (Siegler 1999, 70–73.)

Pusateri ja Thong (2001) tutkivat myöhemmin sanakohtaisten n-best-listojen tuottamista. He huomasivat, että vaikka tunnistin ei palauttaisikaan oikeaa sanaa tunnistuksen ensimmäisenä vaihtoehtona, oikea sana on usein viiden parhaan arvauksen joukossa. He laskivat tunnistustulokselle viiden parhaan sanaan perustuvan tunnistustarkkuuden (5-best accuracy). Verrattuna tunnistustarkkuuteen kun huomioon otettiin pelkästään tunnistimen paras arvaus, tunnistustarkkuus parani viisi parasta arvausta huomioon ottaen noin 5% (Pusateri & Thong 2001).

Dokumenttien laajentaminen

N-best listojen käyttämisessä on se ongelma, että tunnistin ei tee eroa kahden semantiikaltaan erilaisen sanan välillä. Siksi pelkkien n-best-listojen tarkastelu saattaa tuottaaa oikeinkin tunnistettujen sanojen lisäksi täysin aiheeseen kuulumattomia sanoja. Singhal ym. (1998) vertasivat puhetunnistimen tuottamia transkriptioita uutistekstejä sisältävään tietokantaan. Jos tietokannasta löytyi dokumentti, joka sanatasolla oli puheentunnistimen tuottaman transkription kaltainen, dokumenttia verrattiin vielä tunnistimen n-best-listoihin. Tämän jälkeen tekstidokumentista poimittiin n-best-listoilla esiintyvät sanat alkuperäisen tunnistustuloksen rinnalle. Tällainen *dokumentin laajentaminen* (document expansion) lisäsi oikeiden sanojen määrää tunnistustuloksessa, jonka seurauksena myös hakutulokset paranivat. (Singhal ym. 1998.)

Hieman toisenlaista lähestymistapaa sovelsivat Jones ym. (1996). Sen sijaan, että he olisivat laajentaneet dokumentin sanoja tekstikorpuksesta, he yhdistivät kahden eri tunnistimen tuottaman tunnistustuloksen konkatenoimalla tunnistinten tuottamat transkriptiot. Kun haku kohdistettiin yhdistettyyn transkriptioon, hakutulos parani verrattuna yksittäisiin transkriptioihin kohdistuneisiin hakuihin. Tämä johtuu siitä, että toisessa transkriptiossa virheellisenä tunnistettu sana saattoi esiintyä oikein tunnistettuna toisessa. Siten yksittäisten väärin tunnistettujen sanojen vaikutus väheni samalla kuin järjestelmällisesti oikein tunnistetut sanat korostuivat. Myöhemmin myös Sanderson ja Crestani saavuttivat hyviä tuloksia samankaltaista menetelmää käyttämällä (1998).

Ng (2000) teki kokeita, jossa hän yhdisteli kahdeksan eri tunnistimen tuottamia transkriptioita. Tunnistimien sanavirhetasot vaihtelivat 24,6% ja 66,0% välillä. Hän huomasi, että kaikkien kahdeksan tunnistimen transkriptioiden yhdistäminen johti parempaan hakutulokseen kuin mihin pelkästään parasta transkriptiota käyttämällä päästiin. Lisäksi kolmen heikoimman yli 40% sanavirhetason tunnistimen transkriptioiden poistaminen paransi hakutulosta entisestään. (Ng 2000.)

Kyselyjen laajentaminen

Samoin kuin tekstitiedonhaussa myös puhedokumenttien haussa voi kyselyjen laajentamisella helpottaa tiedonhakijan tiedontarpeen ilmaisemista. tehtävää kun hän ilmaisee tiedon tarvettaan. Lisäystermejä käyttämällä voidaan edistää sellaistenkin dokumenttien löytymistä, jossa hakijaa kiinnostavasta aiheesta on puhuttu hakusanoista poikkeavilla sanoilla. Jourlin ym. (1999) kokeilivat erilaisia kyselyjen laajentamistapoja ja totesivat, että tekstitiedonhausta tutut menetelmät parantavat hakutuloksia myös puhehaun yhteydessä.

Kyselyn laajentaminen saattaa osaltaan myös helpottaa virheellisestä transkriptioista johtuvaa tiedonhaun ongelmaa. Laajentamisella parannetaans sen todennäköisyyttä, että ainakin osa kyselyn sanoista esiintyy dokumenttien transkriptioissa oikein tunnistettuna.

Kyselyjen laajentamisessa voidaan lisäksi hyödyntää tietoa tunnistimen tuottamista virheistä hakuavaimiin liitteyn. Jos esimerkiksi tiedetään, että tietyt tai tietyntyyppiset sanat usein tunnistetaan väärin, voidaan hakusanalistaan lisätä myös varsinaisten hakusanojen todennäköiset väärintulkinnat. Chen ym. (2001) sovelsivat yllä esitetyn kaltaista menetelmää, jolloin järjestelmän keskimääräinen tarkkuus parani muutamalla prosentilla. Srinivasan ja Petkovic (2000) puolestaan raportoivat oman järjestelmänsä parantaneen saantia eli löydettyjen dokumenttien määrää. Parannus tapahtui kuitenkin hakutuloksen tarkkuuden kustannuksella, jolloin hakutulokseen tuli myös enemmän hyödyttömiä dokumentteja.

5.2.3 Leksikon rajoittuneisuus

Jos puheentunnistukseen liittyvä yleinen epätarkkuus hetkellisesti unohdetaan, sanakirjapohjaiseen puheentunnistukseen perustuvan tiedonhaun ensisijainen ongelma on leksikon asettamat rajoitteet haettavalle tiedolle. Koska vain leksikossa

esiintyvät sanat pystytään tunnistamaan, merkitsee se sitä, että tunnistettavassa materiaalissa esiintyvät leksikon ulkopuoliset tunnistetaan väistämättä väärin. Tämä tarkoittaa tiedonhaun kannalta sitä, että leksikon ulkopuoliset sanat ja niihin liittyvä informaatiosisältö jäävät hyödyntämättä. Tämä on erityisen ongelmallista, jos leksikossa esiintymättömät ja tunnistamatta jäävät sanat ovat informaatiosisällöltään hakijan kannalta tärkeitä.

Useat tiedonhakuun käytetyistä tunnistimista käsittävät suuria sanavarastoja. Esimerkiksi Sheffieldin yliopiston Abbot-tunnistin tuntee noin 60 000 sanaa (Robinson, Hochberg & Renals 1996) ja LIMSI:n tunnistin 65 000 (Gauvain, Lamel & Adda 2000). Näiden tunnistinten sanavarastot on yleensä valittu tunnistamaan tiettyä materiaalia. Esimerkiksi Allan ja muiden TREC-hakuja varten (Allan ym. 1998) käyttämän tunnistimen 56 000 sanan sanasto valittiin saman aikajänteen tekstiuutisista.

Sanakirjoista jätetään yleensä pois ensisijaisesti harvinaisempia sanoja, koska niiden poisjättämisestä seuraa vähiten virhetunnistuksia. Harvinaiset sanat puolestaan ovat yleensä tiedonhaussa merkityksellisimpiä. Tutkiessaan TREC-konferenssin vuoden 1998 tunnistinten materiaalia ja niihin kohdistettujen hakujen tuloksia, Garofolo ym. laskivat materiaalille kaksi eri sanavirhemäärää: kaikista väärintulkituista sanoista laskettu SWER (Story Word Error Rate) sekä väärin tulkituista henkilöiden, paikkojen ja yritysten nimistä laskettu NEWER (Named Entity WER). He huomasivat, että NEWER korreloi SWER:iä voimakkaammin hakutuloksen laadun kanssa. Erisnimet vaikuttaisivat siten olevan varsin tärkeitä tiedonhaussa ja hyvien hakutulosten tuottaminen edellyttää niiden tunnistamisen puheesta. (Garofolo ym. 1998.)

Lisäksi kieli muuttuu ja varsinkin ajankohtaisia asioita käsittelevä terminologia saattaa olla tiedonhaun kannalta merkityksellinen. Tunnistinten sanavarastoa voidaan tietysti laajentaa. Tällöin on huomioitava se, että tunnistinten sanavarastot ovat jo nyt varsin suuria ja entistä suurempi sanavarasto voi heikentää tunnistustulosta. Ratkaisu voisi olla, että tunnistimen sanavarasto muuttuisi ajan mukaan, laajentumisen sijaan. Muutos voisi tapahtua esimerkiksi automaattisesti saman aikajänteen samoja aihetta käsitteleviä tekstejä tarkkailemalla.

Toisaalta sanakirjaan kuulumattomia (ja tämän takia virheellisesti muiksi sanoiksi tunnistettuja) sanoja voidaan käsitellä samalla tavalla kuin virhetulkintojakin. Singhal ym. (1998) laajensivat tunnistimen tuottamia transkriptioita ensin tunnistimen

n-best listojen perusteella. Myöhemmin he huomasivat, että lisäystermien rajoittaminen pelkästään tunnistimen n-best listoilla esiintyviin sanoihin ei ollut tarpeen. Dokumentin laajentamisen ansiosta hakutulokset paranivat huomattavasti. Dokumenttien laajennus myös vähensi eri puheentunnistinten käytöstä johtuvia hakutulosten välisiä laatueroja. (Singhal & Pereira 1999.)

Sanakirjaan kuulumattomat sanat ovat kuitenkin suuri ongelma silloin, kun ne esiintyvät hakutermeinä. Witbrock ja Hauptmann (1997a; 1997b) tutkivat leksikon ulkopuolisten (out-of-vocabulary, OOV) sanojen vaikutusta haun onnistumiseen. He huomasivat, että yli puolet puhehaun tulosten heikkenemisestä tekstihakuun verrattuna johtui OOV sanoista. Tämän valossa voisi olettaa, että edellä kuvaillut sanakirjapohjaiseen tunnistukseen perustuvan puhehaun menestystarinat perustuvat pitkälti siihen, että tunnistinten sanavarastoja on voitu optimoida sisältämään tunnistettavan materiaalin kannalta keskeisiä sanoja.

5.2.4 Puheen prosodisen informaation hyödyntäminen

Puhesignaali sisältää yksittäisten sanojen äänteiden lisäksi paljon *prosodista* informaatiota, kuten äänenpainoja ja taukoja. Näitä viestinnän välineitä ihmiset hyödyntävät jokapäiväisessä elämässään selventämään ja painottamaan puheensa tietosisältöä. Puheentunnistimet käsittelevät kuitenkin yleensä niin lyhyitä aikaikkunoita, että puheen hitaita prosodisia ominaisuuksia ei saada talteen (Crestani 2001).

Puheen prosodiset ominaisuudet voivat kertoa paljon puheen sisällöstä ilman tietoa siinä käytetyistä sanoista. Esimerkiksi saksankielistä puhetta käsittelevässä EVAR-järjestelmässä pystyttiin painotuksen ja sävelkulun perusteella määrittämään muun muassa oliko puhuttu lause kysymys vai väitelause (Nöth 1991, 133–139). Koska prosodisia keinoja käytetään viestin selkeyttämiseen ja jäsentämiseen ihmistenvälisessä viestinnässä, prosodisesta informaatiosta voi olla hyötyä myös puhetiedonhaussa.

Prosodisen informaation käyttämistä tiedonhaussa ovat tutkineet esimerkiksi Chen ym. (2001). He liittivät tunnistimen tuottamaan transkriptioon tiedon yksittäisten sanojen painoista ja äänenkorkeuden muutoksista. He olettivat, että dokumentissa olevat tärkeät sanat lausutaan hitaammin ja selvemmin painottaen kuin vähemmän tärkeät sanat. Kun hakutuloksien laskemisessa otettiin huomioon

dokumenttien sanojen prosodiset ominaisuudet hakutulokset paranivat hieman. (Chen ym. 2001.) Myös Crestani (2001) on osoittanut, että puheessa käytettyjen painotusten avulla voidaan erottaa semanttisesti keskeisiä sanoja dokumentissa. Lähestymistavat ovat kuitenkin olleet hyvin suoraviivaisia. Esimerkiksi on oletettu, että suurempi paino aina indikoi sisällöllisesti merkittävää sanaa. Puheessa ei kuitenkaan aina painoteta pelkästään tiedonvälityksen kannalta tärkeitä sanoja, vaan usein prosodisia keinoja käytetään korostamaan kokonaisia lauseita (Hansson 2000). Prosodia, kuten muutkin kielen piirteet, ovat myös kielikohtaisia. Iivonen (2000) on tutkinut suomalaisten kysymysten intonaatiota ja toteaa, että suomessa painotuksen ja sävelkulun muutokset esiintyvät vain satunnaisesti.

5.3 Äännetunnistus ja osittaistäsmäytys

Vaihtoehto sanakirjapohjaiselle puhehaulle on tunnistaa audiomuotoisesta materiaalista pelkästään siinä esiintyviä äännejonoja. Äänteisiin pohjautuvan puhehaun ajatuksena on, että puhetta ei tarvitse saattaa tekstimuotoiseksi, jotta puheeseen sisältyvää informaatiota voidaan tallettaa ja käsitellä. Tällöin oletetaan, että äänteidenkin perusteella voidaan löytää se audiotiedosto tai puheen kohta, jossa tietty ilmaisu esiintyy. Tällöin kaikki puheessa esiintyvä informaatio, eli siinä käytetyt sanat leksikosta riippumatta, saadaan talteen. Haun yhteydessä hakusanat esitetään äänneasussaan esimerkiksi prosodista sanakirjaa apuna käyttäen. Tämän jälkeen puhemateriaalista tuotetusta transkriptiosta voidaan etsiä ne kohdat, joissa esiintyy hakusanan kaltainen ääniasu. Jos tällainen löytyy, on mahdollista, että hakusana on puhuttu kyseisessä puhetiedoston kohdassa.

Äännetunnistimen käytön on toivottu ratkaisevan puhemateriaalin haun ongelmia varsinkin sellaisissa hakutilanteissa, joissa hakusanan ei voi olettaa esiintyvän sanakirjoissa. Lisäksi se saattaa olla ainoa menetelmä käsitellä jatkuvasti kasvavia dokumenttimääriä riittävän nopeasti. Laajan sanaston avulla tehtävä puheentunnistus on nimittäin aikaa vievä prosessi. Esimerkiksi IBM:n 64 000 sanan tunnistimelta menee 30-kertainen aika puheaineiston pituuteen verrattuna sen tunnistamiseen (Dharanipragada ym. 1998). Yksittäisten äänteiden tunnistus puolestaan on huomattavasti sanakirjapohjaista tunnistusta nopeampaa (ks. esim. Smeaton ym. 1998).

Ihannetapauksessa puhemateriaalin äänteet tunnistetaan täysin oikein ja puhuja on lisäksi ääntänyt sanan yleisten ääntämysperiaatteiden mukaan. Usein tunnistin ei kuitenkaan tuota täydellistä kuvausta puhemateriaalin äänneasusta, vaan tunnistustuloksessa on virheitä. Lisäksi äännetunnistuksessa erottuu myös yksittäiset ääntämyserot puheessa. Koska yksittäisten äänteiden tunnistuksessa ei myöskään voida hyödyntää tietoa kielen mahdollisista sanoista (joka sanapohjaisessa tunnistuksessa yleensä auttaa rajaamaan pois vaihtoehtoisia tulkintoja), äänteistä koostuva transkriptio sisältää paljon virheitä. Parhaimmillaankin äännetunnistuksessa on 30% virheellisesti tunnistettuja äänteitä (Robinson, Hochberg & Renals 1994). Äännetunnistuksen virheet ovat väärin tunnistettuja äänteitä, eivät siis kokonaisia sanoja kuten sanastopohjaisessa tunnistuksessa.

Äännetunnistin ei myöskään tuota tunnistustulokseen sanarajoja, kuten sanakirjapohjaisessa tunnistuksessa tehdään. Puhtaasti akustisen tiedon perusteella sanarajojen määrittäminen on mahdollista vain, jos puheessa pidetään tauko jokaisen sanan välillä. Näin ei useinkaan luonnollisessa puheessa tehdä, joten transkriptio koostuu puhetta kuvaavista pitkistä äännejonoista.

Äännetunnistukseen perustuvia menetelmiä on sovellettu puhehakuun jo ennen kuin sanojen tunnistamiseen soveltuvia järjestelmiä oli käytössä. Ensimmäisten joukossa olivat Wechsler ja Schäuble (Wechsler & Schäuble 1995; Schäuble & Wechsler 1995), jotka tutkivat puhehakua saksankielisestä uutismateriaalista. Myöhemmin samanlaista lähestymistapaa on käytetty irlantilaisten radiouutisten (Smeaton ym. 1998) sekä TREC-testeissä käytettyyn englanninkielisen uutismateriaalin haussa (Ng & Zobel 1998).

5.3.1 Äänne vai foneemi?

Luvussa 3.2 mainitsin, että termit foneemi ja äänne toisinaan sekoitetaan keskenään ja että jotkut foneemitunnistimen nimikkeellä kulkevat tunnistimet itse asiassa ovat äännetunnistimia. Äänteiden tai toisaalta foneemien käyttö tunnistimessa voi kuitenkin vaikuttaa siihen, miten hakujärjestelmä toimii. Vaikka sanan merkitys ei muuttuisikaan kun jossakin kohtaa sanaa käytetään saman foneemin eri varianttia (ks. luku 2.4), tietty äänneasu voi systemaattisesti liittyä tiettyyn äänneympäristöön. Tällaisten systemaattisten piirteiden taltiointi voi olla haun kannalta tärkeää. Toisaalta äännetunnistuksessa saattaa helpommin nousta esiin myös sekä erot eri puhujien että saman puhujan eri puhetapahtumien välillä.

Tiedonhaun kannalta merkityksellistä lienee se, että hakija usein on kiinnostunut myös niistä dokumenteista, joissa hakusana lausutaan hieman normaalista poikkeavalla tavalla. Tällöin hakijalle ei ole merkitystä sillä, mitä foneemin varianttia jossakin sanan kohdassa on käytetty. Äänteiden taltiointi saattaa siksi toimia tiedonhaun tavoitteiden vastaisesti ja korostaa tiedonhaun kannalta merkityksettömiä piirteitä. Toisaalta, jos hakija on kiinnostunut esimerkiksi jollakin tietyllä murteella tuotetusta puheesta, hienojakoisemmasta tunnistuksesta voi olla hyötyä.

Ng ja Zue (1997) vertasivat äänteisiin sekä äänteistä muodostettujen laajempien kategorioiden eroja tiedonhaun onnistumiseen. He huomasivat, että laajemmat foneettiset kategoriat palvelevat parhaiten tiedonhaun tarkoitusta. Tutkijoiden soveltamista laajemmista kategorioista yksikään ei täysin vastannut englannin kielen foneemijaottelua. Tulokset antavat kuitenkin perusteet olettaa, että foneemeihin perustuva karkeampi jaottelu on äänteisiin perustuvaa jaottelua parempi vaihto tiedonhaun kannalta. (Ng & Zue 1997.)

5.3.2 Äännetunnistetun puheen indeksoinnista

Tehokkaan tiedonhaun kannalta on tärkeää, että dokumentit voidaan tutkia nopeasti ja tämän jälkeen esittää tulokset tiedonhakujärjestelmän käyttäjälle. Tekstitiedonhaussa, kuten sanakirjapohjaisia tunnistusmenetelmiä hyödyntävässä puhehaussakin, on tapana muodostaa haettavasta aineistosta käänteistiedosto eli *indeksi*. Indeksi sisältää jokaiseen tietokannassa esiintyvään (tiedonhaun kannalta merkitykselliseen) sanaan liitettyjä tietoja siitä, missä tietokannan dokumenteissa sana esiintyy ja kuinka monta kertaa. Kun järjestelmä suorittaa varsinaisen haun, dokumentteja ei enää tarvitse erikseen tutkia, vaan hakutulos tuotetaan indeksiä tutkimalla. (Ks. esim. Järvelin 1995, 96–105.)

Pelkällä äännetunnistuksella ei pystytä erottelemaan puheesta sanoja, joten äännetunnistukseen ei suoraan voi käyttää perinteisiä käänteistiedostoja. Jotta puhetta voitaisiin indeksoida käänteistiedostojen avulla, puhetta kuvaavia äännejonoja pitää ensin pilkkoa.

Puhedokumenttien indeksointi n-grammien avulla

Yksi lähestymistapa tehokkaan puhetiedonhaun takaamiseksi on ollut pilkkoa puhedokumenttien äännejonoja vakiomittaisiin, osittain päällekkäisiin, merkkijonoihin eli niin kutsuttuihin *n-grammeihin* (esim. Wechsler & Schäuble 1995; Ng & Zue 1997; Wechsler 1998; Smeaton ym. 1998; Ng & Zobel 1998). N-grammit muodostetaan siten, että grammattavasta merkkijonosta muodostetaan kaikki n merkin mittaiset, peräkkäisistä merkeistä koostuvat merkkijonot. Tekstin merkeistä muodostetuissa n-grammeissa on lisäksi usein käytössä tyhjä merkki kuvaamassa sanarajaa. Tekstin n-grammauksessa ei myöskään yleensä anneta yhden merkkijonon jatkua sanasta toiseen. Tällöin n-grammeissa on tyhjää joko alussa tai lopussa, mutta ei ikinä keskellä. Kuviossa 5 on muodostettu 1-, 2- ja 3-grammit äännejonosta [foneemi] sekä tyhjiä merkkeinä käyttämällä 1-, 2- ja 3-grammit sanasta foneemi (tyhjänä merkkinä käytetty *).

[f] [o] [n] [e] [e] [m] [i]	f o n e e m i
[fo] [on] [ne] [ee] [em] [mi]	*f fo on ne ee em mi i*
[fon] [one] [nee] [eem] [emi]	**f *fo fon one nee eem emi mi* i**

Kuvio 5. 1-, 2- ja 3-grammit äännesekvenssistä [foneemi] sekä tekstisanasta foneemi.

Kun dokumentti on esitetty tietokannassa n-mittaisten osamerkkijonojen joukkona, sen käsittelyyn voi soveltaa monia tekstitiedonhaun menetelmiä. Vakiomittaisiin merkkijonoihin perustuvan tiedonhaun suorituskyky riippuu siitä, miten hyvin indeksoinnissa käytetyt n-grammit kuvaavat alkuperäisiä merkkijonoja. On selvää, että yksittäisten merkkien sisältämä informaatio on sen verran pieni, ettei niiden avulla pääse riittävään erottelukykyyn. Tuloksellisen puhehaun kannalta ei merkittävästi ole hyötyä siitä tiedosta, missä kaikissa dokumenteissa on esiintynyt äänne [e]. Toisaalta hyvin pitkien n-grammien käyttäminen indeksissä johtaa nopeasti siihen, että indeksin tutkiminen hidastuu, samalla kuin indeksissä on monta esiintymää saman sanan eri vaihtoehdoista. Puhehaussa sanan eri vaihtoehdot syntyvät pitkälti tunnistusvirheistä. Esimerkiksi äännejonot [sullmenn], [sullmemn] ja [ksulmemn] ovat kaikki mahdollisia äännetunnistimen antamia tunnistustuloksia lausahduksesta ”suomen”. Myös sanan taivutusmuodot voivat aiheuttaa sen, että

hakijaa kiinnostava sana esiintyy indeksissä useassa eri muodossa, joten indeksin täyttymisen ongelma liittyy myös suomenkieliseen tekstihakuun (vrt. luku 4.2).

Tiedonhaun onnistumisen kannalta on tärkeää löytää sellaisia puhetta kuvaavia äännejonoja, jotka ovat tarpeeksi pitkiä jotta ne sisältävät tiedonhaun kannalta tärkeää informaatiota ja muodostaa n-grammit näistä. Mitä pidempiä n-grammit ovat, sen enemmän ne sisältävät juuri lähdedokumentilleen ominaista informaatiota. Toisaalta pidempiin grammeihin mahtuu myös enemmän tunnistusvirheitä. Tämän takia pidempien grammien käsittelyssä korostuu puheentunnistuksen epätäydellisyydestä aiheutuvat tunnistusvirheet.

Ng ja Zue (1997) vertasivat erimittaisten osien soveltuvuutta englanninkielisessä puhehaussa. He totesivat että 3-grammit antoivat parhaimmat hakutulokset. Lyhyemmissä grammeissa oli liikaa toistoa, ts. 2-grammit eivät riittämissä määrin kuvanneet dokumentin informaatiosisältöä. Pitkissä grammeissa puolestaan tunnistusvirheet aiheuttivat sen, että kaksi samaa sanaa eivät juuri koskaan sisältäneet samoja grammeja. (Ng & Zue 1997.)

N-grammien avulla muodostettuihin indekseihin perustuva puhehaku on osoittautunut toimivaksi ratkaisuksi sellaisissa tilanteissa, joissa sanakirjapohjaista tunnistinta ei syystä tai toisesta ole voitu käyttää. N-grammeihin perustuvaa ja sanakirjapohjaista puhehakua vertailevat tutkimustulokset osoittavat kuitenkin, että keskimäärin sanakirjapohjaiset menetelmät suoriutuvat n-grammeihin perustuvia menetelmiä paremmin (ks. esim. Ng, Wilkinson & Zobel 2000). N-grammien etu on kuitenkin se, etteivät ne sido haun mahdollisuuksia tietyn leksikon sanoihin. Ng (2000) onkin osoittanut, että n-grammien käyttö yhdessä sanakirjapohjaisen menetelmän kanssa voi parantaa hakutuloksia verrattuna tuloksiin, kun käytetään jompaa kumpaa menetelmää yksinään.

Yksittäisten sanojen löytämiseen perustuvat järjestelmät

Äännetunnistetun puheen indeksointi perustuen vakiomittaisiin merkkijonoihin on ollut hyvin yleinen lähestymistapa nopean puhetiedonhaun takaamiseksi. Toinen lähestymistapa on ollut yksittäisten sanojen löytämiseen perustuvien (*word spotting*) menetelmien käyttäminen. Näissä menetelmissä puhetta ei pilkota ennen käsittelyä, vaan hakusanojen esiintymiä etsitään tutkimalla kokonaisia äännejonoja

puhedokumenteissa. (Ks. esim. James 1995; James 1996; Wechsler 1998; Ferrieux & Peillon 1999; Amir, Efrat & Srinivasan 2001.)

Word spotting -menetelmää soveltavissa järjestelmässä itse tiedonhaku voidaan toteuttaa pitkälti samalla tavalla kuin sanakirjapohjaiseen tunnistukseen perustuvissa järjestelmissä. Ero edellisiin on siinä, että nyt puhemateriaalin tunnistus tapahtuu yksi sana kerrallaan. Sanakirjapohjaisissa järjestelmissä järjestelmän tuntemat sanat määritellään ennen tunnistamisprosessin käynnistämistä. Koska tunnistusprosesseja on vain yksi, järjestelmälle tunnistuksen aikana vieraat sanat jäävät lopullisesti puuttumaan tunnistustuloksesta, vaikka ne esiintyisivätkin itse dokumenteissa. Word spotting -menetelmät perustuvat siihen, että yksittäisen (tai muutaman) sanan tunnistusprosessi voidaan käynnistää aina uudelleen sen jälkeen, kun hakija on ilmaissut, mistä sanoista hän on kiinnostunut.

Sellaisten dokumenttien käsittely, joita ei ole indeksoitu, vie luonnollisesti huomattavan paljon enemmän aikaa kuin indeksoitujen dokumenttien käsittely. Prosessi hidastuu entisestään, jos tunnistustulokseen sisällytetään yhden ainoan tulkinnan lisäksi tunnistimen vaihtoehtoisia tunnistustuloksia. James (1995) käytti puhedokumenttien äänteiden kuvaamiseen tunnistimen tuottamia vaihtoehtoisia tulkintoja, jotka esitettiin suunnatun graafin avulla (ns. äännelattiisi). Aluksi lattiisien läpikäyminen vei suhteettoman paljon aikaa. Myöhemmin hän kehitti menetelmän, jolla hakusanan skannaus kaikkiaan 2 tuntia 27 minuuttia pitkästä puhemateriaalista kesti 3,2 sekuntia. Toisin sanoen järjestelmä operoi noin kahdeksantuhatkertaisella nopeudella normaaliin puhenopeuteen nähden (James 1996). Ferrieux ja Peillon (1999) kehittämässä järjestelmässä puolestaan keskipituisen (pituus 4 tavua) kyselyn vertaaminen tietokannan puhedokumentteihin tapahtui keskimäärin 100 000 äänteen sekuntivauhdilla. Raportoidut nopeudet riippuvat luonnollisesti hakutilanteessa käytetyistä tietokoneista. Tietokoneiden prosessointitehot ovat kasvaneet vuosi vuodelta, mutta niin ovat myös kasvaneet tietokantojen koot. Yllä esitetyt prosessointinopeudet johtavat todennäköisesti pitkiin odotusaikoihin, jos menetelmiä käytetään sellaisenaan kokonaisien tietokantojen käsittelyyn.

Yksi mahdollisuus nopeuttaa käsittelyä on lisätä word spotting -menetelmien avulla löydetyt sanat (tai sanojen todennäköiset esiintymät) indeksiin joka kerta kun löydetään uusia sanoja. Näin ensimmäisen word spotting soveltamiskerran jälkeen ei samalle sanalle tarvitsekaan enää käyttää aikaavieviä algoritmeja, vaan sana voidaan etsiä indeksistä. Indeksi tällöin kasvaa asteittain käyttäjien ilmaistessa järjestelmälle,

mistä sanoista he ovat kiinnostuneita. Esimerkiksi Jamesin edellä mainittu äännelattiisijärjestelmä yhdistettiin myöhemmin Video Mail Retrieval -järjestelmään. Järjestelmä säilyttää kerran word spotting -menetelmän avulla löydetyt sanat ja käyttää näitä yhdessä sanakirjapohjaisesta tunnistustuloksesta muodostetun indeksin kanssa (Brown ym. 1996).

5.3.3 Äännejonojen osittaistäsmäytys

Tiedonhaun onnistumisen kannalta on olennaista, että hakujärjestelmä pystyy löytämään myös virheellisesti tunnistettuja, mutta puheessa esiintyneitä sanoja. Virheellisesti tunnistetutkin sanat voidaan löytää, jos hakusanan täsmäytyksessä käytetään osittaistäsmäytykseen perustuvia menetelmiä. Tällöin materiaalista etsitään hakusanan äänneasun kaltaisia, mutta ei välttämättä aivan identtisiä kohtia. Jos dokumentin transkriptiosta löytyy riittävän lähellä hakusanaa oleva äänneasu, oletetaan, että haettava sana on puhuttu dokumentissa.

Erilaisia osittaistäsmäytysmenetelmiä on paljon ja niiden erityispiirteet riippuvat pitkälti siitä, millaisten merkkijonojen vertailuun ne on kehitetty. Kattavan selvityksen erilaisista menetelmistä ja niiden käyttötavoista antavat esim. Hall ja Dowling (1980), Navarro (2001) tai Hyyrö (2000). Seuraavaksi tarkastelen kahta äännetunnistukseen perustuvassa puhehaussa sovellettua menetelmää nimittäin editointietäisyyttä ja n-grammitäsmäytystä.

Editointietäisyyteen perustuva äänteiden osittaistäsmäytys

Suuri osa puhehaussa käytetyistä menetelmistä perustuvat niin kutsutun *editointietäisyyden* laskemiseen. Yksinkertaisimmillaan kahden merkkijonon välinen editointietäisyys on pienin mahdollinen määrä vaihto-, lisäys- ja poisto-operaatioita, joita tarvitaan toisen merkkijonon tuottamiseksi ensimmäisestä merkkijonosta.

Editointietäisyyttä laskiessa on mahdollista määritellä jotkin operaatiot kalliimmiksi kuin toiset. Esimerkiksi määrittelemällä poisto- ja lisäysoperaatiot kalliimmiksi kuin vaihto-operaatio voidaan vaikuttaa siihen, että järjestelmä antaa pienempiä arvoja sellaiselle operaatiosekvenssille, joka tuottaa samanmittaisia merkkijonoja. Puhe puolestaan sisältää erilaisia äänteitä ja osa äänteistä sekoittuvat tunnistuksessa helpommin keskenään kuin toiset.

Tunnistustuloksia tarkastelemalla voidaan jokaisen tunnistimessa määritellyn äänteen kohdalla selvittää, millä todennäköisyydellä se tunnistetaan väärin joksikin toiseksi äänteeksi. Tämä tieto voidaan sisällyttää editointietäisyyden laskemiskaavaan. Muokkaamalla tällä tavalla editointietäisyysarvoa vaikutetaan siihen, että suurella todennäköisyydellä keskenään sekoittuvat äänteet eivät vaikuta sanojen samankaltaisuuteen yhtä paljon kuin harvinaisemmat vaihdokset. Tällaista menetelmää (tosin hieman toisistaan eroavin laskutavoin) ovat puhehaussa soveltaneet muun muassa Wechsler (1998) sekä Amir, Efrat ja Srinivasan (2001).

N-grammit virhesietoisena täsmäytysmenetelmänä

Puheen indeksoinnissa suositut n-grammit toimivat sellaisenaan virhesietoisena täsmäytysmenetelmänä. Tämä lisäetu tulee, kun puhedokumentteja pilkotaan osamerkkijonoiksi. Nimittäin kun äännejonoa kuvataan tietyllä joukolla n-grammeja, kahden äännejonon samankaltaisuutta voidaan laskea tutkimalla näiden yhteisien n-grammien määrä.

Erilaisia n-grammimenetelmiä on kauan hyödynnetty virheitä sisältävän materiaalin täsmäytykseen tekstitiedonhaussa. Esimerkiksi automaattisen tekstintunnistuksen (Optic Character Recognition, OCR) avulla tuotetut dokumentit sisältävät usein väärintunnistettuja merkkejä, jotka vaikeuttavat hakua. Kirjoitusvirheiden käsittelyssä n-grammeja on käytetty sekä virheiden löytämiseen että niiden korjaamiseen. N-grammien käytöllä voi myös vähentää sekä kyselyn että dokumentin sisältämistä kirjoitusvirheistä johtuvia täsmäytysongelmia. Samoin niitä on käytetty helpottamaan tiedonhakua historiallisista teksteistä, joissa ongelmana ovat kielen kehityksen johdosta muuttuneet sanat (Robertson & Willett 1992). N-grammit ovat myös osoittautuneet erinomaisen hyödyllisiksi silloin, kun materiaalissa ei ole ollut helposti määriteltäviä sanarajoja. Tämä pitää paikkansa esimerkiksi monissa aasian kielissä Esimerkkeinä mainittakoon Kiinan (Chen, A. ym. 1997; Nie ym. 2000), Japanin (Fujii & Croft 1993) ja Korean (Lee & Ahn 1996) kielet.

N-grammit toimivat sellaisenaan osittaistäsmäytysmenetelmänä. Tästä huolimatta osittaistäsmäytystä on myös sovellettu puheen indeksoinnissa käytettyjen n-grammien keskinäiseen täsmäytykseen. Erityisesti lyhyemmillä n-grammeilla tulokset ovat kuitenkin olleet huonoja, koska vaikuttaisi siltä, että osittaistäsmäytys huonontaa entuudestaan lyhyiden merkkijonojen erottelukykyä (Ng & Zobel 1998).

6 Kokeita suomenkielisessä puhehaussa

Suomenkielisestä puhehausta ei ole tehty aiemmin kokeellista tutkimusta. Tällä hetkellä ei myöskään ole olemassa järjestelmiä, joiden avulla suomenkielistä puhetta voisi hakea sisällön perusteella. Siksi kehitin Suomenkielisen puhemateriaalin haku -projektissa järjestelmän, joka mahdollistaa suomenkielisen puhehaun ja samalla tarjoaa testialustan erilaisten puhehaun menetelmien tutkimukselle.

Suomenkielistä tiedonhakuun soveltuvaa ja laajan sanavarastoon perustuvaa sanakirjapohjaista tunnistinta ei vielä ole olemassa. Tällä hetkellä suomenkielisillä puheentunnistusjärjestelmillä on mahdollista suorittaa vain hyvin suppeaan sanakirjaan pohjautuvaa puheentunnistusta. (Ks. luku 3.4.) Tämän takia puhehaussa ylivoimaiseksi lähestymistavaksi todettu sanakirjapohjainen tunnistus ei ole vaihtoehto tässä tutkimuksessa. Sen sijaan minulle tarjoutui mahdollisuus käyttää Tampereen teknillisessä yliopistossa kehitettyä äännetunnistinta. Siksi päätin kehittää äännetunnistukseen perustuvan hakujärjestelmän.

6.1 Puhedokumenttien suodattaminen n-grammien avulla

Kuten luvussa 5.3 todettiin, äänneperusteisessa puhehakuun on ehdotettu kahta päälähestymistapaa: n-grammeihin perustuva puheen indeksointi ja toisaalta word spotting -tekniikat, jotka operoivat pitkillä äännejonoilla. Molemmissa menetelmissä on hyvät ja toisaalta huonot puolensa.

N-grammit ovat nopea ja suoraviivainen tapa käsitellä puhetta. Yksinkertaisesta lähtökohdastaan huolimatta n-grammit sisältävät riittävästi informaatiota puhedokumenttien hakemiseen. Lisäksi puhedokumentteja voi pilkkoa n-grammeiksi jo ennen hakutilannetta, jolloin hakutuloksen tuottaminen on varsin nopeaa. N-grammien muodostuksessa ei kuitenkaan oteta huomioon puheen sisältöä, vaan äännejono jaetaan osiin systemaattisesti sisällöstä riippumatta. N-grammeilla ei myöskään päästä yhtä hyviin tarkkuuksiin, kuin mihin word spotting -menetelmillä ylletään.

Word-spotting -menetelmillä voidaan monipuolisesti verrata hakusanaa puhedokumenttien sisältämiin äännejonoihin. Lisäksi word spotting mahdollistaa vaihtoehtoisten tunnistustulosten huomioonottamisen ja prosodisen informaation hyödyntämisen. Hienostuneempien menetelmien käyttäminen on samalla kuitenkin äärimmäisen aikaavievää ja äännejonojen käsittely voidaan aloittaa vasta hakutilanteessa, kun tiedetään mitä dokumenteista haetaan.

N-grammit mahdollistavat puheen indeksoinnin ja prosessoinnin ennen varsinaista hakua. Word spotting -algoritmit puolestaan toimivat vasta hakusanan antamisen jälkeen eli hakutilanteessa. Ihannetapauksessa tiedonhakujärjestelmässä yhdistyisi n-grammien suoraviivaisuus ja toisaalta word spotting -algoritmien monipuoliset täsmäytysmenetelmät. Oletukseni oli, että n-grammien avulla voitaisiin vähentää word spotting -algoritmien läpikäymää tietomäärää. Tällöin myös haun toteuttamiseen tarvittava aika pienenee.

Tekstitiedonhakua varten on toisinaan muodostettu dokumenteista niin kutsuttuja *nimikirjoituksia* (signature, text signature). Nämä tiedostot ovat eräänlaisia hash-tauluja varsinaisista haettavista dokumenteista. Niiden avulla kaikista tietokannan dokumenteista muodostetaan vakiomittaisia esityksiä, joissa yksittäisillä biteillä kuvataan dokumenttien informaatiosisältöä. (Ks. Järvelin 1995, 131–133; Ashford & Willett 1988, 89–91) Kun jokaista tietokannan dokumenttia edustaa vakiomittainen ja samaa formaattia noudattava, 0- ja 1-biteistä muodostuva nimikirjoitus, näiden samankaltaisuutta voidaan verrata yksinkertaisilla AND ja OR-operaatioilla. Näin päästään hyvin lyhyihin käsittelyaikoihin dokumenttia kohden.

Nimikirjoitustiedostot voidaan muodostaa muun muassa n-grammien avulla. Tällöin dokumentista tuotetaan ensin tietynmittaisista n-grammeista koostuva joukko, jonka jälkeen nimikirjoitustiedoston jokainen dokumentin n-grammia edustava bitti muutetaan nollasta yhdeksi. (Ks. esim. Robertson & Willett 1998.) Dokumenttien samankaltaisuutta kuvaa niissä olevien yhteisten n-grammien määrä. Osittaistäsmäytystä käsittelevässä kirjallisuudessa samankaltainen toteutus kulkee nimellä q-gram-suodatus (Hyyrö 2000). Navarro (2001) on kokeellisesti demonstroinut menetelmän nopeutta suhteessa muihin osittaistäsmäytysmenetelmiin.

6.2 Tutkimuskysymykset

Ensisijainen tavoitteeni on tutkia miten hyvin n-grammien avulla muodostetut nimikirjoitustiedostot soveltuvat puhedatan suodattamiseen. Tähän liittyen haluan selvittää seuraavia seikkoja:

- Millä *n* arvoilla n-grammit antavat laadullisesti parhaimman suodatuksen?
- Voiko suodatusta soveltaa kokonaisista uutisista tuotettuihin nimikirjoituksiin? Saavutetaanko pienempiä uutisten osia tutkimalla parempia tuloksia? Jos näin, mikä on sopiva uutisen osa nimikirjoituksen muodostamiselle?
- Mitkä hakusanan ominaisuudet vaikuttavat suodatuksen laatuun ja miten?

Edellisten lisäksi halusin myös selvittävää, missä määrin tunnistimen koulutuksessa sovellettua nyrkkisääntöä ”lausutaan niin kuin kirjoitetaan” voisi hyödyntää tiedonhakujärjestelmissä. Lisäsin kysymyksen:

- Voiko puhedokumenttien suodatuksessa käyttää sekä teksti- että puhemuodossa esitettyjä hakusanoja?

Seuraavaksi esitellään tutkimuksessa käytetty puhemateriaali. Loput tästä luvusta käsittelee tutkimusta ja sen tuloksia. Ensin esitellään tutkimuksessa käytetty puhemateriaali luvussa 6.3. Tämän jälkeen käydään läpi transkription tuottamista ja merkkivirhemäärän optimointia luvussa 6.4. Luvussa 6.5 verrataan eri n-grammien suorituskykyä uutisten suodatukseen. Alaluvussa 6.5.4 selvitetään, miten hakusanan ominaisuudet vaikuttavat suodatuksen laatuun. Menetelmien vaikutusta käsiteltävään dokumenttimäärään arvioidaan alaluvussa 6.5.5. Lopuksi pohditaan suodatuksen merkitystä puhehaun järjestelmissä luvussa 6.6.

6.3 Tutkimuksessa käytetty puhemateriaali

Tutkimusta varten ei ollut saatavilla tiedonhaun tutkimuksen edellyttämää määrää valmista puhemateriaalia. Jotta suodatusmenetelmien kokeellinen vertailu olisi mahdollista, tutkimusta varten piti luoda realistiseen tietokantaan verrattavissa oleva puhetietokanta.

Puhetietokannan uutisiksi valittiin sanomalehtiartikkeleita sisältävästä TUTK-tietokannasta yhteensä 288 artikkelia koskien 17 eri uutisaiheesta. Uutiset koskevat tapahtumia vuosilta 1988–1992 ja niiden joukossa on sekä koti- että ulkomaan uutisia sekä talousuutisia. (Sormunen 2000, 59.) Editoin tekstimuotoiset artikkelit puheuutisten kaltaisiksi, jonka jälkeen luin ne nauhalle kaiuttomissa olosuhteissa TTY:n signaalinkäsittelyn laitoksen audiolaboratoriossa. Tavoitteena oli tuottaa pienimuotoinen tietokanta, jonka sisältämien puheuutisten laatu olisi samanlainen kuin lyhyistä (studio-oloissa tuotetuista) radiouutisista nauhoitettu uutismateriaali. Tämän vuoksi pitkiä tekstiuutisia piti ensin lyhentää. Lyhennys tapahtui poistamalla kokonaisia kappaleita. Lisäksi tekstiuutisista poistettiin ilmaisutapoja, jotka eivät sellaisenaan ole siirrettävissä puheeseen, kuten sulkujen sisälle kirjoitetut tarkennukset ja lisätiedot. Näiden suhteen editointi suoritettiin tapauskohtaisesti. Yleislinjana oli poistaa kaikki ylimääräinen. Näin uutiset saatiin mahdollisimman yksinkertaisiksi ja enemmän radiouutisten kaltaisiksi.

Hakusanoja varten luin 232 erillistä sanaa, joiden tiedettiin esiintyvän uutisissa. Sanojen valintaperuste oli se, että niiden avulla voitaisiin myöhemmin ilmaista kokonaisia kyselyitä kerätyssä uutisaineistossa edustettuina oleville aiheille. Lisäksi kaikki hakusanat ovat vähintään viisi merkkiä pitkiä. Tämä siksi, että n-grammia ei voi muodostaa pienemmästä kuin n merkkiä pitkästä merkkijonosta ja tutkimuksessa haluttiin käyttää n arvoja 2–5. Luin hakusanat kahteen kertaan, jolloin jokaisesta sanasta saatiin kaksi eri tunnistusversiota.

Aineisto on luettu kaiuttomissa oloissa, joten puhetallenteen laatu on hyvä. Puheen taustalla ei ole muita ääniä (esimerkiksi musiikkia). Uutisten puhujalla on hieman kokemusta radiotyöstä ja puheen tyyli on verrattavissa uutispuheeseen.

Puhutut uutiset ovat keskimäärin 93 sanaa pitkiä ja ne kestävät hieman alle minuutin (59,8s). Tiedonhaun tutkimuksen näkökulmasta käytetty aineisto on häviävän pieni (288 uutista, kestoltaan yhteensä 4,5 tuntia). Aineisto on vaatimaton myös suhteessa muualla tehdyssä puhetiedonhaun tutkimuksessa käytettyyn aineistoon (vrt. luku 5.2.1). Aineiston koon takia tutkimuksesta ei voi vetää johtopäätöksiä menetelmien yleisestä suorituskyvystä isommassa tietokannassa. Sen sijaan menetelmien väliseen paremmuusjärjestykseen vaikuttaa ensisijaisesti puhemateriaalin laatu eikä sen määrä. Siksi tutkimuksessa käytetty pienempikin aineisto sallii menetelmien keskinäisen vertailun, johon tässä työssä keskityn.

Puhetallenteet ovat hyvälaatuisia, mikä helpottaa puheentunnistustehtävää. Tästä syystä tuloksista ei voi vetää suoria johtopäätöksiä siitä, miten menetelmät suoriutuvat taustahälyisen puheen hakemiseen. On kuitenkin realistista olettaa, että osa haettavasta puhemateriaalista on tutkimuksen materiaaliin verrattavissa olevaa hyvälaatuista puhetta. Tämän takia on perusteltua tutkia myös hyvänlaatuisen puheen hakemiseen soveltuvia menetelmiä.

6.4 Tunnistin ja transkription muodostaminen

Tutkimusta varten sain käyttööni TTY:lla kehitetyn puheentunnistimen. Tunnistin perustuu TTY tutkijaryhmän Kivimäki, Lahti & Koppinen (2000) raportoimaan kätkettyihin Markovin malleihin perustuvaan tunnistimeen. Tunnistinta on myöhemmin kehittänyt eteenpäin tutkija Timo Pylvänäinen.

Tunnistin on pohjimmiltaan äännetunnistin, johon on myöhemmin liitetty suomenkielisistä osasanoista koostuva osasanaleksikko (vrt. luku 3.1). Tunnistin on siinä mielessä erikoinen, että se on koulutettu tekstimateriaalin avulla. Samoin tunnistimen osasanaleksikko perustuu suomenkielisessä tekstimateriaalissa esiintyvin kirjainsekvensseihin.

Tekstimateriaalin avulla toteutetusta koulutusprosessista seuraa se, että tunnistimen oppimat akustiset mallit liittyvät tekstissä esiintyviin kirjaimiin puheen äänteiden tai foneemien sijasta. Tästä erityispiirteestä johtuen, tunnistimen tuottama tunnistustulos koostuu kirjaimista. Koska alkuperäisessä tekstissä on esiintynyt kirjaimia, jotka eivät vastaa mitään tiettyä äännettä puheessa (esimerkiksi kirjain z), tunnistin tuottaa myös tuloksessaan ei-foneettisia merkkejä. Täten tunnistimen tulosta ei voi varsinaisesti pitää foneettisena transkriptiona.

6.4.1 Perustranskriptio

Tunnistin tuottaa joukon vaihtoehtoisia tulkintoja, joihin on liitetty todennäköisyysarvo 100 ms aikaikkunoittain. Kuviossa 6 on pieni osa tunnistimen tuottamaa tunnistusvektoria. Vektori sisältää joukon aikaikkunoita, joissa jokaisessa on ensimmäisenä aikaikkunan leima, jonka jälkeen seuraa vaihteleva määrä tunnistusvaihtoehtoja. Lisäksi jokaista tunnistusvaihtoehtoa seuraa tunnistimen sille antama todennäköisyysluku. Todennäköisyysarvot ovat logaritmisia, toisin sanoen

-1000.00 tarkoittaa $10^{-1000} = 1/10^{1000}$. Suurempi arvo tarkoittaa kuitenkin aina myös suurempaa todennäköisyyttä, toisin sanoen -400 todennäköisyydellä varustettu merkkijono on todennäköisempi kuin -600:lla.

<4800.00, taise, -1935.3942, ansai, -1641.7653><4900.00, ansai, -1641.7653><5000.00, ansai, -1641.7653, yksi, -1614.1221><5100.00, yksi, -1614.1221><5200.00, yksi, -1614.1221><5300.00, yksi, -1614.1221, m, -405.2286, j, -710.8741> <5400.00, j, -710.8741, lis, -1094.2379><5500.00, lis, -1094.2379><5600.00, lis, -1094.2379, y, -652.3317>

Kuvio 6. Osa tunnistimen tuottamaa tunnistusvektoria.

Ensimmäistä tunnistustulosta varten poimin jokaisesta aikaikkunasta parhaimman todennäköisyyden saaneen tunnistusehdokkaan. Nämä tunnistusehdokkaat liitettiin peräkkäin tunnistustulokseksi, jolloin syntyi BASE-transkriptio (ks. taulukko 1, s.50).

6.4.2 Transkription esikäsittely

Tutkiessani BASE-transkriptiota huomasin, että tunnistustuloksena syntynyt transkriptio järjestelmällisesti oli alkuperäistä uutistekstiä pidempi. Tämä johtuu siitä, että tunnistin tuottaa tunnistusehdokkaita jokaista 100ms ikkunaa kohden. Toisinaan tunnistin tuottaa aikaikkunaa kohden vain yhden merkin, mutta usein tunnistimen arvaukset ovat monesta merkistä koostuvia sekvenssejä. 100ms on kuitenkin niin lyhyt aika, ettei siinä ehdi sanomaan kovin monta äännettä. Tästä syntyi ajatus jättää osa tunnistustuloksessa olevista aikaikkunoista pois lopullisesta transkriptiosta.

Toista transkriptiota varten poimin tunnistustuloksen peräkkäisistä aikaikkunoista saatavista samoista merkkijonoista vain toisen. Esimerkiksi jos kahdessa peräkkäisessä ikkunassa oli merkkijono 'ta' sellaisella todennäköisyydellä, että ne tulisivat transkriptioon, vain toinen näistä otettiin mukaan. Tästä versiosta käytetään nimitystä PICK-transkriptio.

Kokeilin myös poimia tunnistustuloksia aikaikkunoista transkriptioon muilla kuin pelkän todennäköisyytensä perusteella. Oletin, että pitkät merkkijonot sisältävät enemmän tiedonhaun kannalta hyödyllistä informaatiota kuin lyhyemmät merkkijonot. Pidemmät merkkijonot saavat kuitenkin yleensä heikompia todennäköisyysarvoja kuin lyhyemmät merkkijonot. Järjestin siksi yhden aikaikkunan sisältämät tunnistustulokset paremmuusjärjestykseen keskenään siten, että jokaiselle merkkijonolle laskettiin arvo jakamalla sen saama oikeintunnistamista kuvaava todennäköisyysarvo tunnistustuloksen merkkien määrällä. Tämän jälkeen poimin

suuremman arvon saaneen tunnistustuloksen mukaan transkriptioon. Käytän tästä transkriptiosta nimitystä SELECT.

SELECT-menetelmä suosii pitkiä merkkijonoja, mutta ottaa samalla huomioon niiden todennäköisyydet. Lopulliseen transkriptioon otettiin edelleen peräkkäisistä ikkunoista saatavista useista samoista tunnistustuloksista vain yksi.

Minulla ei ollut tietoa siitä, miten hyvin tunnistin oli koulutettu käsittelemään kaksoiskonsonantteja tai -vokaaleja. Oletin, että tunnistin saattaisi tuottaa sellaisia peräkkäisiä aikaikkunoita, jotka yhdistettynä tuottavat peräkkäisiä merkkejä kuvaamaan äänteitä, joita puheessa kuitenkin on esiintynyt vain lyhyesti. Tämän takia kokeilin poistaa kaikki duplikaattimerkit tavuja konkatenoimalla tuotetusta transkriptioista. Näin syntyivät viimeiset verrattavat menetelmät PICK-STRIP ja SELECT-STRIP.

6.4.3 Tunnistuksen merkkivirhetaso

Merkkijonojen poimimismenetelmien kehittämisvaiheessa minulla oli käytössäni 98 yksittäisen lauseen puheaineisto. Tälle piloottiaineistolle ei suoritettu merkkivirhearviointia, mutta se toimi apunani, kun kehitin erilaisia tunnistustuloksen esikäsittelymenetelmiä. Lopullisen materiaalin tunnistuksessa ilmeni, että tunnistinta oli muokattu piloottiaineiston tunnistamisen jälkeen, ennen lopullisen aineiston tunnistamista. Tunnistusvektoreiden tarkastelusta huomasin, että lopullinen tunnistin vaikuttaisi itsessään jo suosivan pitkiä merkkijonoja, jolloin SELECT-menetelmän merkitys vähenee. Lisäksi muokattuun tunnistimeen oli liitetty uusi osasanaleksikko, joka myös osaltaan voi vaikuttaa sen toimintaan.

Seuraavaksi esittelen eri menetelmillä (BASE, PICK, SELECT, PICK-STRIP ja SELECT-STRIP) saatuja merkkivirhetasoja. Tunnistimen luonteesta johtuen päätin arvioida sen tuottaman transkription virhetasoa siten, että vertasin transkriptiota alkuperäisiin teksteihin (joista ensin oli poistettu kaikki erikoismerkit sekä välilyönnit sanojen väliltä). Virhetaso lasketaan korvaus-, poisto- ja lisäysoperaatioiden minimaalisen määrän prosentuaalisena osuutena alkuperäisen tekstin merkkimäärästä. On korostettava, että luku voi nousta yli 100 prosentin jos tarkastelun kohteena oleva transkriptio on pitempi kuin alkuperäisteksti.

Edellä kuvaillulla tavalla laskien BASE-transkription keskimääräiseksi merkkivirhetasoksi saatiin 112%. PICK–menetelmän tuottama transkriptio saavutti keskimääräisen merkkivirhetason 42,5 %. SELECT saavutti merkkivirhetason 43,6%.

Duplikaattikirjainten poisto paransi merkkivirhetasoa molemmilla menetelmillä, PICK-STRIP sai merkkivirhetason 42,0% ja SELECT-STRIP puolestaan 42,6%.

Paras keskimääräinen merkkivirhetaso suhteessa alkuperäistekstiin on 42,0%, joka saavutettiin PICK-STRIP-menetelmällä. Alla olevassa taulukossa on esitetty erilaisin menetelmin tuotetut transkriptiot, alkuperäisteksti sekä menetelmien keskimääräiset merkkivirhetasot.

Alkuperäisteksti: yhdysvaltain ehdotuksesta kertoivat nimettöminä pysyttelevät kokouslähteet		
Menetelmä	**Transkriptio**	**Merkki-virhetaso**
BASE	yyityisityisvaltvaltainainenäätäättuksästästästerkerkerkpaivavamimehtehteminemineminerityrityteylevtelevatkatkukukuusuuslalahtäätäät	112%
PICK	yityisvaltainenäättuksästerkpaivamimehtemineritytelevatkukuuslahtäät	42,5%
SELECT	yityisvaltainenäättuksästerkpaivaimehtemineritytelevatkukuuslaäät	43,6%
PICK-STRIP	yityisvaltainenätuksästerkpaivamimehtemineritytelevatkukuslahtät	42,0%
SELECT-STRIP	yityisvaltainenätuksästerkpaivaimehtemineritytelevatkukuslaät	42,6%

Taulukko 1. Transkription esikäsittelymenetelmät. Taulukossa jokaisen menetelmän kohdalta esimerkki transkriptiosta sekä menetelmän avulla saavutettu keskiarvoinen merkkivirhetaso.

Puhehaun empiirisiä kokeita varten otin käyttöön matalimman merkkivirhetason saaneen PICK-STRIP-transkription. Samoin käsittelin puhutuista hakukysymyksistä saadut tunnistusvektorit PICK-STRIP-menetelmällä.

6.5 Puhedokumenttien suodattaminen nimikirjoitustiedostojen avulla

Tutkimuksen tarpeisiin toteutin Java-ohjelmointikielellä tiedonhaun testausympäristön, jonka avulla voi järjestää tietokannan sisältämiä puhetiedostoja

paremmuusjärjestykseen luvussa 6.1 kuvailtujen nimikirjoitusten avulla. Järjestelmällä voi hakea puhetiedostoja sekä puhutun että tekstimuotoisena annetun hakusanan avulla. Järjestelmän parametreja muuttamalla voidaan muunnella suodatuksessa käytettyjä menetelmiä.

Järjestelmä käy läpi kaikki tietokannan sisältämistä puheuutisista muodostetut nimikirjoitukset ja luo näiden perusteella tuloslistan. Tuloslistassa dokumentit esitetään suodatusmenetelmän antaman vertailuarvon mukaisessa paremmuusjärjestyksessä. Mikäli useampi dokumentti saa saman vertailuarvon, dokumentit listataan tulokseen siinä järjestyksessä, jossa ne esiintyvät tietokannassa.

Yksittäisten nimikirjoitusten samankaltaisuuden laskemiseen käytetään OVERLAP-kaavaa (Salton & McGill 1983, 203–204):[6]

$$\mathrm{Sim}(a, b) = \frac{|A \cap B|}{\min\{|A|, |B|\}}$$

Kaavassa a ja b ovat tutkittavat merkkijonot ja A ja B ovat näiden nimikirjoitusten kuvaamat grammijoukot. Toisin sanoen A on tutkittavasta dokumentista ja B käytetystä hakusanasta n-gram-tekniikalla saatujen osamerkkijonojen joukko.

6.5.1 Tulosten arviointiperusteet

Tiedonhakujärjestelmän arvioinnissa on tärkeää määritellä, onko sen löytämät dokumentit *relevantteja* vai ei. Toisin sanoen: ovatko sen kyselyihin palauttamat dokumentit niitä, joita järjestelmän olisi pitänyt palauttaa. Yleisesti ottaen relevanssi on epämääräinen käsite ja toisinaan voi olla hankala arvioida, onko jokin dokumentti tiedon hakijalle hyödyllinen tai oikea. Tässä tutkimuksessa tarkastellaan *teknistä relevanssia*, jolloin relevanteiksi katsotaan kaikki ne puhedokumentit, joissa hakusana on esiintynyt alkuperäisessä puheessa. Sana saa esiintyä puheessa joko sellaisenaan tai taivutetussa muodossa ja joko yhdyssanan osana tai yksinään. Tulosten arvioinnissa käytetään *binääristä relevanssiarviota*, jolloin jokainen tuloksessa esiintyvä dokumentti on joko relevantti tai sitten ei – mitään välimuotoja ei ole.

Tutkimuksessa käytetyistä hakusanoista tiedetään, että tietokanta sisältää jokaista hakusanaa kohden vähintään yhden relevantin dokumentin. Keskimäärin

[6] Toinen vaihtoehto samankaltaisuuden laskemiselle olisi ollut DICE-kaava $|A \cap B| / |A \cup B|$ (Salton & McGill 1983, 203–204). DICE kuitenkin soveltuu lähinnä yksittäisten sanojen vertailuun, koska se eriarvoistaa eripituisia verrattavia merkkijonoja.

jokaista hakusanaa kohden on 8,6 relevanttia dokumenttia. Luku vaihtelee yhdestä peräti 63 relevanttiin dokumenttiin.

Tutkimuksessa keskityttiin yksittäisten sanojen avulla suoritettavaan suodatukseen. Suodatusmenetelmän laatua arvioidaan järjestelmän tuottaman tuloksen perusteella. *Tarkkuus* ilmaisee kuinka suuri osa järjestelmän löytämistä dokumenteista ovat relevantteja. *Saanti* puolestaan ilmaisee kuinka suuri osa koko tietokannan sisältämistä relevanteista dokumenteista hakumenetelmän avulla löydettiin. Tässä tutkimuksessa kaikki dokumentit kirjoitetaan hakutuloslistaan, joten lopullisen tuloksen saanti on aina 100%. Seuraamalla saannin kehittymistä voidaan kuitenkin nähdä miten tarkkuus vaihtelee eri *saantitasojen* mukaan. Esimerkiksi 10% saantitaso tarkoittaa, että relevanteista dokumenteista on löydetty yksi kymmenesosa.

Menetelmien keskinäiseen vertailuun käytetään *keskimääräistä tarkkuutta* (average precision) yli saantitasojen. Luku kuvaa sitä, kuinka paljon menetelmä keskimäärin kykenee löytämään relevantteja dokumentteja suhteessa ei-toivottuihin dokumentteihin. Toisena arviointiperusteena käytetään tarkkuutta 100 prosentin saantitasolla. Tästä luvusta käytetään jatkossa nimitystä *lopullinen tarkkuus*. Lopullinen tarkkuus kuvaa, kuinka suuri osuus relevantteja dokumentteja on suhteessa ei-toivottuihin dokumentteihin siinä vaiheessa kun kaikki tietokannan sisältämät relevantit dokumentit on listattu tulokseen.

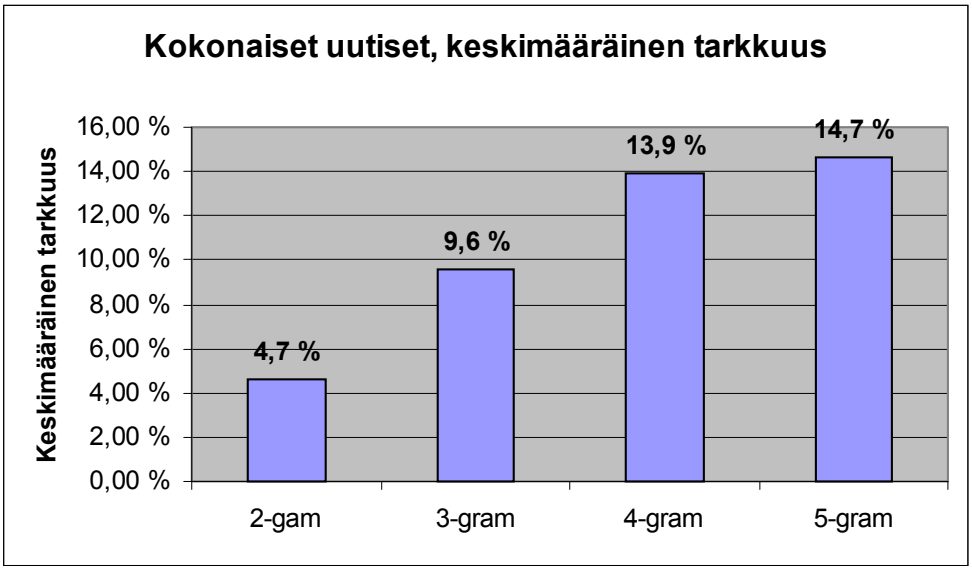
Menetelmien tilastollisessa vertailussa on käytetty Conoverin esittämää versiota Friedmanin testistä (Conover 1980, 299). Kyseessä on ei-parametrinen testi, jota yleisesti on käytetty tiedonhaussa menetelmien vertailuun. Friedmanin kaltaiset rankiin perustuvat testit soveltuvat käytettäväksi silloin, kun vertailtavat arvot eivät noudata normaalijakaumaa. Tiedonhaussa yksittäisten hakujen vaikutus keskimääräiseen tarkkuuteen on yleensä suurempi kuin menetelmien vaikutus, joten parametristen menetelmien käyttö ei ole perusteltua. (Hull 1993.)

6.5.2 Kokonaisista uutisista muodostetut nimikirjoitukset

Eri grammikokojen vertailua varten muodostin ensin nimikirjoituksia kokonaisista puhedokumenteista. Tätä varten kokonaiset puhedokumenttien transkriptiot pilkottiin erikokoisiin n-grammeihin (n arvoilla 2, 3, 4 ja 5). Tämän jälkeen jokaisesta dokumentista muodostettiin nimikirjoitukset jokaista grammikokoa varten. Suodatuksessa käytettiin puhuttuja hakusanoja. Yksittäisten hakusanojen

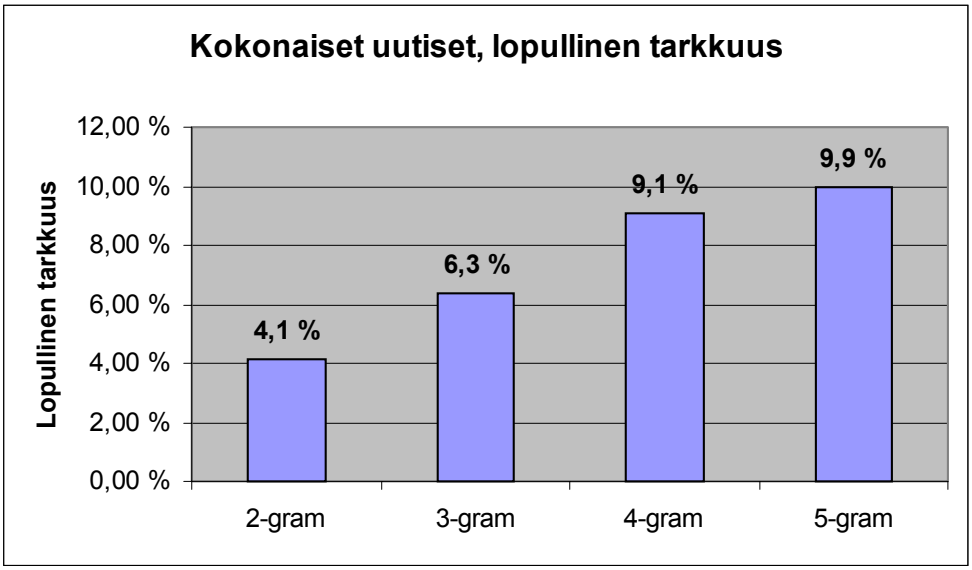
transkriptioista muodostettiin uutisten tavoin 2, 3, 4 ja 5-grammeihin perustuvat nimikirjoitukset. Dokumenttien ja hakusanojen samankaltaisuus laskettiin Overlap-kaavalla ja näin saatuja arvoja käytettiin suoraan tuloksen muodostamiseen.

Kuviossa 7 on esitetty menetelmien suoritus keskimääräisinä tarkkuuksina. Suodatuskyky on kaikilla menetelmillä kohtuullisen heikko. Kaikkein heikoin se on 2-grammeilla, joiden saavuttama tarkkuus on vain 4,7%. Parhaimpaan suoritukseen päästään käyttämällä pitkiä n-grammeja, mutta niissäkin tarkkuus jää alle 15% (5-grammeilla 14,7%). Tämä tarkoittaa, että keskimäärin jokaista relevanttia dokumenttia kohden suodatuksen läpäisee myös kuusi epärelevanttia dokumenttia.



Kuvio 7. Keskimääräiset tarkkuudet, kun suodatuksessa käytettävä nimikirjoitus muodostetaan kokonaisista uutisista.

Kuviossa 8 on esitetty järjestelmän lopulliset tarkkuudet, eli tarkkuus kun saanti on 100%. 2-grammeilla saavutetaan heikoin lopullinen tarkkuus 4,1%. Parhaimmat arvot saadaan, mitä pidempiä n-grammeja käytetään; 4-grammeilla 9,1% ja 5-grammeilla 9,9%.



Kuvio 8. Lopulliset tarkkuudet, kun suodatuksessa käytettävä nimikirjoitus muodostetaan kokonaisista uutisista.

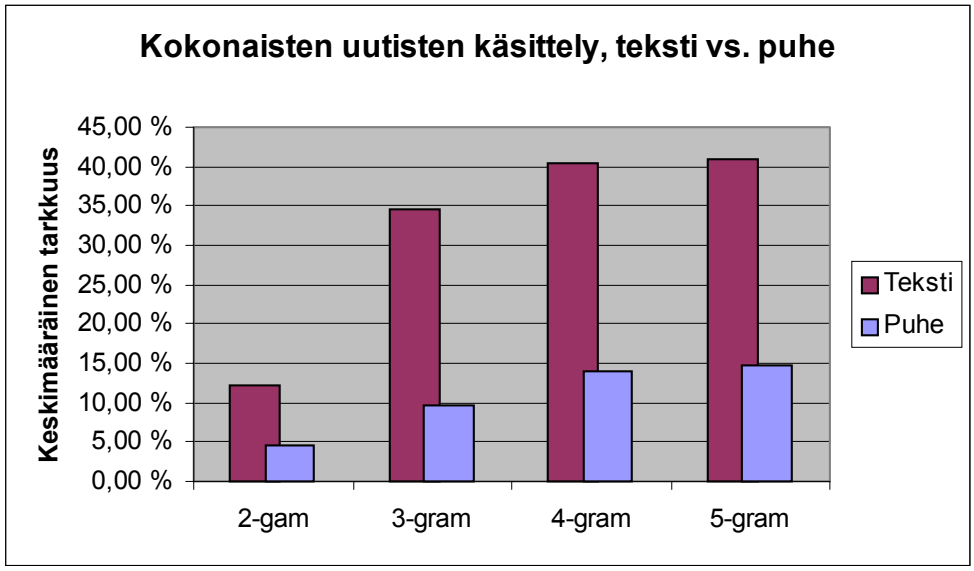
Suodatusmenetelmän tehokkuuteen voi vaikuttaa negatiivisesti kaksi seikkaa. Ensinnäkin tulokseen vaikuttaa puheentunnistuksen laatu. Jos puheentunnistin on tehnyt tunnistusvirheen, joko hakusanan tai puhedokumentin transkriptiossa esiintyvä merkkisekvenssi ei kuvaa alkuperäisessä puheessa esiintyneitä äänteitä.

Toiseksi voi olla, että nimikirjoitusten muodostaminen kokonaisista dokumenteista kadottaa liikaa informaatiota. Nimikirjoitusten muodostuksessa kadotetaan tieto osamerkkijonojen sijainnista ja mahdollisten useampien esiintymien lukumäärästä, koska nimikirjoitukseen merkitään vain tieto siitä, että tietty osamerkkijono on esiintynyt dokumentissa. Tällöin siitä, että dokumentin nimikirjoitus sisältää samoja merkkijonoja kuin hakusanan nimikirjoitus, ei vielä voi päätellä että hakusana on puhuttu dokumentissa. Kielessä on käytössä rajattu määrä erilaisia mahdollisia merkkisekvenssejä, joita käytetään sanojen osina. Mitä pidemmistä dokumenteista nimikirjoitukset muodostetaan, sen todennäköisempää on että jotkut tietyt osamerkkijonot esiintyvät useammassa eri kohdassa ja nimikirjoitukseen syntyy päällekkäisyyttä.

Käsiteltävät uutiset olivat keskimäärin noin minuutin mittaisia. Suodatuksessa käytetyt transkriptiot ovat keskimäärin 712 merkkiä pitkiä. Käsiteltävien dokumenttien pituus lisää päällekkäisten merkkijonojen todennäköisyyttä, jolloin myös nimikirjoitusten erotteluvoima heikkenee. Toisin sanoen pitkien dokumenttien

kohdalla nimikirjoitus alkaa täyttymään kielen yleisistä merkkijonoista, eikä enää ensisijaisesti kuvaa dokumentin tietosisältöä.

Selvittääkseni johtuiko suodatuskyvyn heikko tulos nimikirjoitusten liiasta päällekkäisyydestä, suoritin n-grammiperusteisen suodatuksen myös tekstikysymyksillä tekstidokumenteista. Dokumentteina käytettiin puheuutisten täydellisiä transkriptioita, joista poistettiin sanavälit. Sanavälit poistettiin, jotta materiaalia voisi verrata aitoihin transkriptioihin suoritettuihin suodatuksiin.



Kuvio 9. Keskimääräiset tarkkuudet tekstillä tekstistä sekä puhehaulla.

Kuviossa 9 on esitetty edellä puheesta saadut tulokset suhteessa ihannetranskriptiosta saatuihin. Tulokset ovat järjestelmällisesti parempia kuin puheella. Kahden mittaiset osamerkkijonot suoriutuvat edelleen heikosti. Parhaiten suoriutuvat pitkät n-grammit, mutta nyt ero 4- ja 5-grammien välillä on jo varsin pieni (0,5%).

On kuitenkin huomattava, että suodatusmenetelmä ei ihanteellisellakaan transkriptiolla juurikaan ylitä yli 40% tarkkuutta (5-gram: 41,0%). Tämä viittaisi siihen, että pitkistä dokumenteista muodostetut nimikirjoitukset täyttyvät liiaksi päällekkäisistä osamerkkijonoista. Seuraavaksi tarkastellaan dokumenttien suodatusta perustuen nimikirjoituksiin, jotka on tehty pienemmistä puhedokumenttien osista.

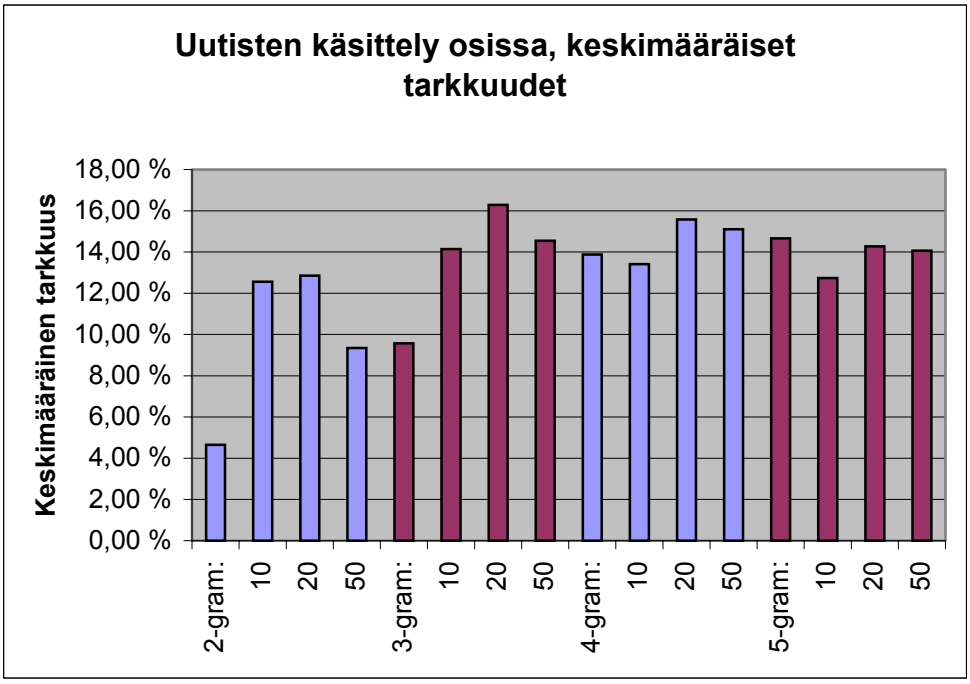
6.5.3 Uutisten käsittely osissa

Yksi mahdollinen tapa ratkaista päällekkäisten grammien ongelmaa on käsitellä uutisia pienemmissä osissa. Seuraavaa koeasetelmaa varten uutisten transkriptioita

pilkotaan ennen nimikirjoitusten muodostusta osittain päällekkäisiksi merkkijonoiksi (seuraava merkkijono alkaa aina edeltävän merkkijonon puolivälistä). Käytetyt merkkijonojen pituudet olivat 10, 20 ja 50 merkkiä. Jotta tämä käsittely erottuisi grammauksesta käytän siitä jatkossa nimitystä ikkunointi. Ikkunoiksi jakamisen jälkeen jokaisesta uutisikkunasta muodostetaan nimikirjoitukset samoin kuin edellisessä kokeessa. Hakukysymyksinä käytetään puhuttuja hakusanoja. Hakusanoista ei muodosteta ikkunoita, vaan jokaisesta hakusanasta muodostetaan yksi nimikirjoitus.

Tulos tuotetaan siten, että ensin verrataan hakusanan nimikirjoitusta ikkunakohtaiseen nimikirjoitukseen Overlap-kaavalla. Tämän jälkeen uutiselle annetaan vertailuarvo joka on maksimi sen ikkunoiden saamista vertailuarvoista. Tätä vertailuarvoa käytetään, kun uutiset järjestetään paremmuusjärjestykseen tuloksen muodostuksessa. Kuten edellä, saman vertailuarvon saaneet kaksi tai useampaa dokumenttia esitetään tuloksessa tietokannan sisäisessä järjestyksessä.

Kuviossa 10 on esitetty keskimääräiset tarkkuudet eri ikkunakooille. Vertailun helpottamiseksi on vieressä esitetty myös kokonaisilla uutisilla saadut suodatustarkkuudet. Tulokset vahvistavat oletuksen, että pienempien grammikokojen heikompi suodatuskyky johtuu liian pitkistä kerralla käsiteltävistä merkkijonoista.



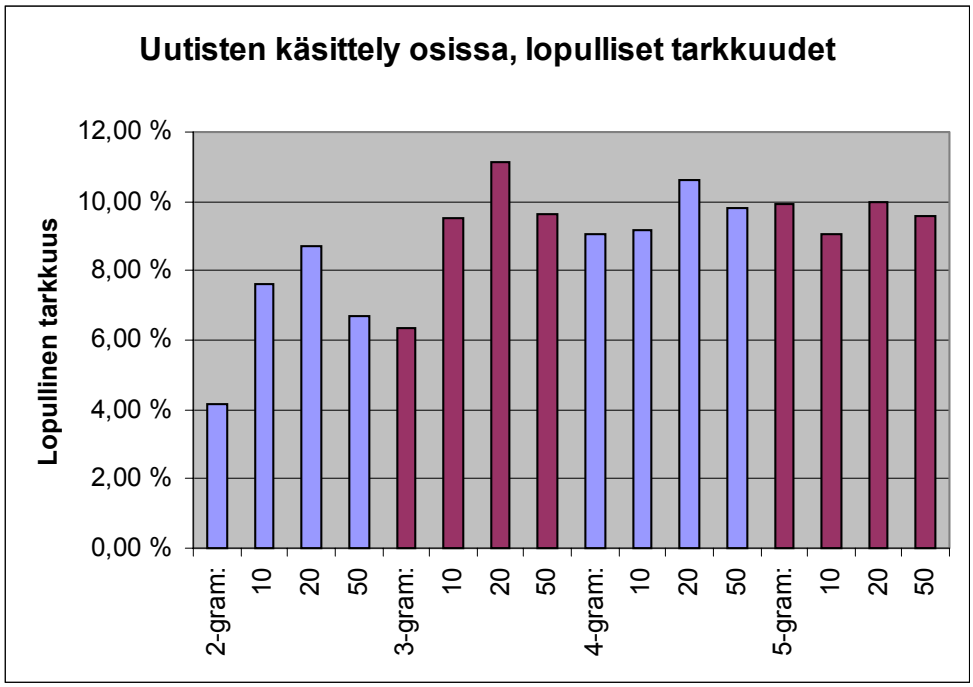
Kuvio 10. Keskimääräiset tarkkuudet eri ikkunakooille.

Varsinkin pienemmillä grammeilla 2-gram ja 3-gram ikkunointi parantaa tuloksia olennaisesti. Ero on tilastollisesti merkitsevä 2-grammeilla kaikilla ja 3-grammeilla 20 merkin ikkunan kooilla. 4- ja 5-grammien suoritukseen ikkunoinnilla ei ole tilastollisesti merkitsevää vaikutusta.

Pienemmillä grammeilla on tärkeää, että käytetty ikkunakoko on suhteellisen pieni. 2-grammeilla huomataan että jo 50 merkin ikkunakoko tuo nimikirjoitukseen liikaa päällekkäisyyttä, jolloin tulokset heikkenevät. Pidemmillä grammeilla ei tätä ongelmaa ole. Esimerkiksi 5-grammeilla ikkunointi vaikuttaisi itse asiassa heikentävän hakutulosta.

Myös liian pieni ikkunakoko voi haitata suodatusta. Tämä johtuu siitä että 41% hakusanoista on yli 10 merkin mittaisia. Koska uutiselle laskettu vertailuarvo on maksimi sen ikkunoiden vertailuarvoista, tämä tarkoittaa sitä että tulos saattaa heiketä kun hakusana on jakautunut usealle ikkunalle. Tässä tapauksessa vain osa hakusanasta mahtuu kerrallaan käsiteltävään 10 merkin ikkunaan. Jos juuri tässä hakusanan osassa on paljon tunnistusvirheitä, ikkunan saama arvo on matala. Samalla paremmin tunnistettu hakusanan osa saattaa täyttää vain puolet viereisestä ikkunasta, jolloin ikkunan muu sisältö heikentää sen saamaa vertailuarvoa. Lopputulos on, että kaksi peräkkäistä ikkunaa antavat molemmat keskinkertaisen vertailuarvon, joista jompikumpi päätyy uutisen vertailuarvoksi.

Ikkunointia käyttämällä myös grammien paremmuusjärjestys muuttuu. Paras suodatustulos 16,3% (keskimääräinen tarkkuus) saadaan 3-grammeilla 20 merkin ikkunalla. 4-grammit eivät myöskään jää kauas kärjestä; 20 merkin ikkunalla päästään 15,6% keskimääräiseen tarkkuuteen. Mielenkiintoista on huomata, että myös 2-grammien avulla voidaan ikkunoita käyttäessä saavuttaa varsin hyviä tarkkuuksia. 20 merkin ikkunaa käyttäen 2-grammeilla saavutetaan kilpailukykyinen 12,9% keskimääräinen tarkkuus verrattuna kokonaisia uutisia käsittelemällä saatuun arvoon 4,7%.



Kuvio 11. Lopulliset tarkkuudet eri ikkunakooille.

Kuviossa 11 on esitetty eri ikkunakooilla saavutetut lopulliset tarkkuudet. Kuviosta huomaa, että parhaimman menetelmän (3-gram, 20 merkin ikkuna) lopullinen tarkkuus on keskimäärin 11,1%. Tämä tarkoittaa sitä, että suodatuksen läpäisee jokaista relevanttia dokumenttia kohden myös yhdeksän epärelevanttia dokumenttia.

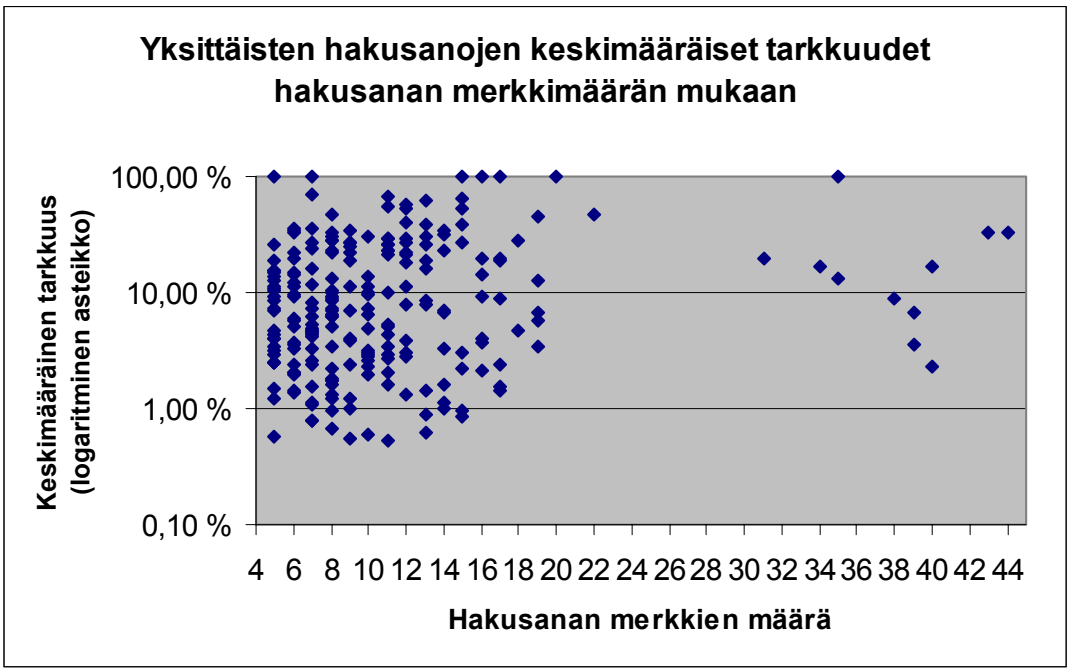
6.5.4 Kokeita hakusanoilla

Edellä on esitetty joukko tuloksia koskien erilaisten suodatusmenetelmien suorituskykyä, kun hakukysymyksenä on käytetty yksittäisiä puhuttuja sanoja. Järjestelmän käytännön soveltamisen kannalta on mielenkiintoista selvittää tarkemmin, hakusanojen vaikutus suodatukseen. Samalla halusin selvittää, johtavatko jotkut hakusanan ominaisuudet järjestelmällisesti keskivertoa parempaan tai huonompaan suodatustulokseen. Vaikutuksia tarkastellaan sekä yksittäisten hakusanojen suorituksista laskettuina keskiarvoina, mutta ennen kaikkea mielenkiintoista on tutkia järjestelmän toimintaa yksittäisillä hakusanoilla. Vaikka keskiarvot antavat hyvän yleiskuvan menetelmän suorituskyvystä, ne voivat johtaa vääriin käsityksiin menetelmän toiminnasta yksittäisillä hakusanoilla.

Hakusanan pituuden vaikutus suodatukseen

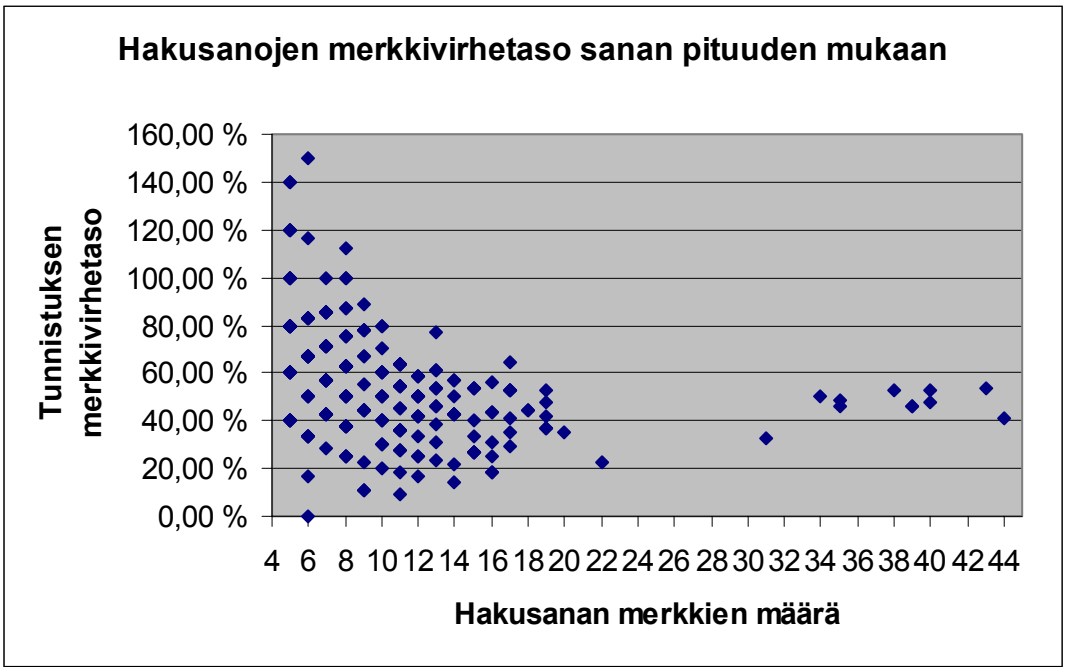
Puheentunnistuksen luonteesta johtuen oletukseni on, että pidemmillä hakusanoilla päästään parempiin tuloksiin kuin käytettäessä lyhyempiä hakusanoja. Tämä olettamus perustuu siihen, että pidemmissä hakusanoissa on suuremmalla todennäköisyydellä enemmän suodatuksen kannalta olennaisia merkkejä kuin lyhyemmissä sanoissa. Vaikka oikeintunnistuksen aste pysyykin samana erimittaisilla sanoilla, oikeintunnistettujen merkkien absoluuttinen määrä kasvaa myös sanan pituuden kasvaessa.

Rajoitan yksittäisten hakusanojen tarkastelun tässä työssä pelkästään parhaita keskimääräisiä suorituksia tuottaneen menetelmän (3-gram, 20 merkin ikkunointi) avulla suoritettuihin suodatuksiin.



Kuvio 12. Yksittäisten hakujen keskimääräiset tarkkuudet hakusanan merkkimäärän mukaan järjestettynä (3-gram, 20 merkin ikkunointi).

Kuviossa 12 on esitetty yksittäisten hakusanojen keskimääräiset tarkkuudet sanojen ihannetranskription merkkimäärän mukaan järjestettynä. Kuviosta näkyy, että yksittäisten hakujen suodatus vaihtelee alle 1% keskimääräisestä tarkkuudesta aina 100% keskimääräiseen tarkkuuteen saakka. Hakusanan pituudella ei kuitenkaan vaikuttaisi olevan säännönmukaista vaikutusta suodatuksen tarkkuuteen.



Kuvio 13. Hakusanan merkkivirhetaso sanan pituuden mukaan.

Kuviossa 13 on esitetty hakusanan merkkivirhetaso sanan merkkien määrän mukaan. Kuviosta näkyy, että varsinkin lyhyillä hakusanoilla on joskus erittäin korkeita merkkivirhetasoja. Erot selittyvät kuitenkin pitkälti tunnistustuloksen ja ihannetranskription pituuksien erosta. Nimittäin sama määrä ylimääräisiä tai puuttuvia kirjaimia aiheuttaa suhteessa suuremman merkkivirhetason pudotuksen lyhyillä hakusanoilla, koska editointietäisyys jaetaan oikean merkkijonon merkkimäärällä.

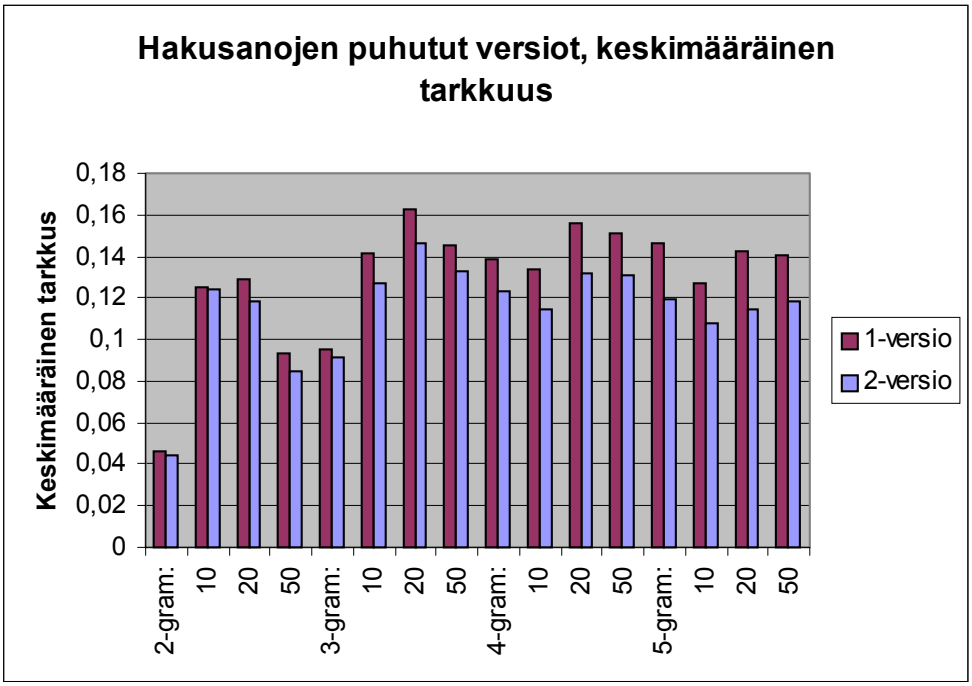
Hakusanan tunnistustarkkuuden vaikutuksesta

Ajatukseni, että pidemmät hakusanat johtaisivat parempiin tuloksiin perustui olettamukseen, että pidemmät hakusanat sisältäisivät enemmän oikeintunnistettuja merkkejä kuin lyhyet hakusanat. Näin ei kuitenkaan näyttäisi välttämättä tapahtuvan. Siksi haluan seuraavaksi selvittää, miten hakusanan tunnistustarkkuus vaikuttaa haun onnistumiseen. Taustalla on halu yleisesti selvittää, miten paljon materiaalin tunnistamiseen käytetty puheentunnistin vaikuttaa suodatuksen onnistumiseen.

Edellisissä luvuissa esitetyt tulokset koskivat suodatusmenetelmien toimintaa yksittäisillä hakusanoilla. Kokeissa käytetyt hakusanat luettiin kahteen kertaan, joista edellä esitetyt tulokset ovat 1-version tuloksia. 1-version keskimääräinen merkkivirhetason on 53,0%. 2-version merkkivirhetaso on 54,3% eli 2,6% huonompi

kuin 1-version. Puhuttuja hakusanoja kuuntelemalla selviää, että 1-version sanat on puhuttu hieman nopeammalla puherytmillä kuin 2-version. On siis mahdollista että puherytmi on vaikuttanut puheentunnistimen tunnistustarkkuuteen.

Selvittääkseni tunnistustarkkuuden vaikutusta suodatuksen laatuun suoritin uutismateriaalin suodatuksen myös keskimääräisesti heikomman tunnistuksen 2-versiota käyttäen. Tämän jälkeen vertasin suodatustuloksia keskenään. Eri puhekertojen menetelmäkohtaiset keskimääräiset tarkkuudet on esitetty kuviossa 14.

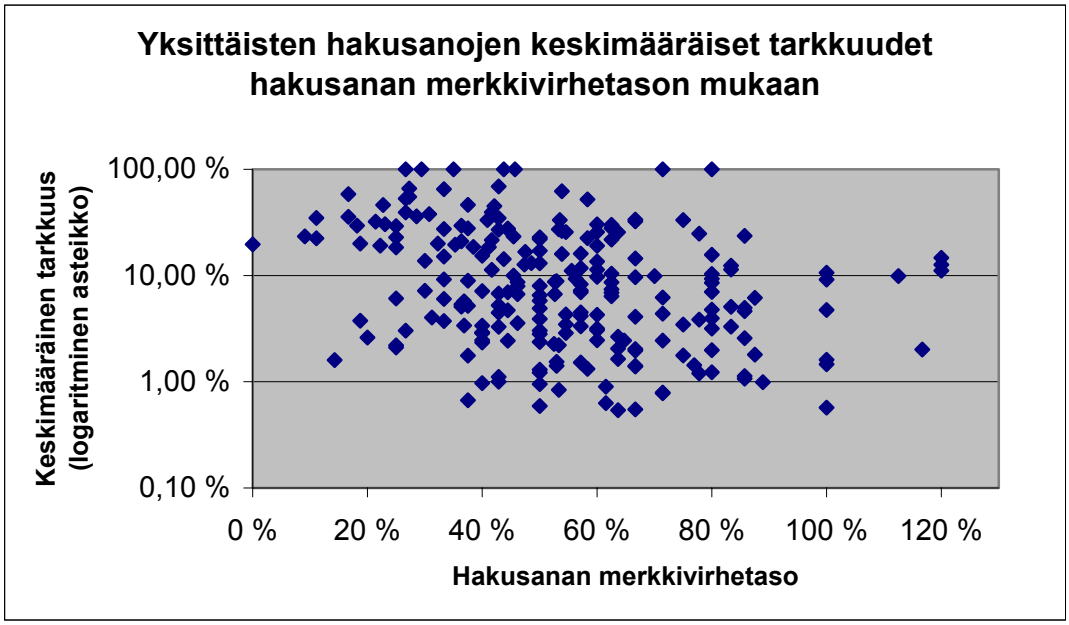


Kuvio 14. Keskimääräiset tarkkuudet eri puhekertojen hakusanoilla.

Kuviota 14 tarkastelemalla huomataan, että 1-version antamat suodatuksen keskimääräiset tarkkuudet ovat järjestelmällisesti hieman parempia kuin 2-version. Pienin ero on n-grammien pienillä *n* arvoilla ja mitä suuremmaksi grammikoko kasvaa, sen suuremmaksi erot kasvavat. Tästä huomataan, että tunnistusvirheet vaikuttavat myös n-grammeihin. Etenkin lyhyempien n-grammien osittaistäsmäytys lieventää tunnistusvirheiden vaikutusta. Hakusanojen merkkivirhetason muutos ei tässä yhteydessä vaikuttanut menetelmien keskinäiseen paremmuusjärjestykseen.

Hakusanojen merkkivirhetason vaikutus haun onnistumiselle

Edellä todettiin, että hakusanojen lukukerran merkkivirhetasolla oli pieni vaikutus haun onnistumiseen. Kuviossa 15 on tarkasteltu yksittäisten hakusanojen merkkivirhetason vaikutusta hakusanan keskimääräiseen tarkkuuteen. Kuviosta voi nähdä jonkinlaista trendiä hakusanan kasvavan merkkivirhetason ja suodatuksen keskimääräisen tarkkuuden heikkenemisen välillä. On kuitenkin merkillepantavaa, että jopa 120% merkkivirhetason hakusana saattaa tuottaa yli 10% keskimääräisiä tarkkuuksia kun taas hyvinkin matala merkkivirhetaso ei aina yllä samaan lukuun. Tässä yhteydessä on muistettava, että puhutulla hakusanalla haettaessa dokumenttien tunnistusvirheet eivät aina tuota ongelmia. Puhutulla hakusanalla haettaessa voi nimittäin olla, että sekä hakusana että sanan esiintymä puheessa on tunnistettu samalla tavalla väärin.



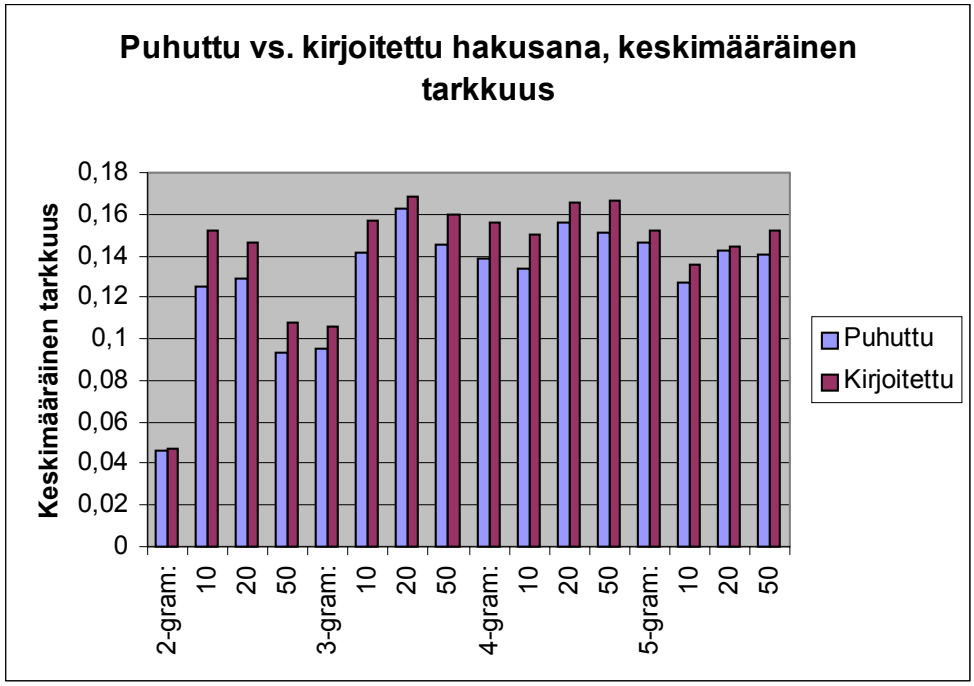
Kuvio 15. Yksittäisten hakujen keskimääräiset tarkkuudet hakusanan merkkivirhetason mukaan järjestettynä (3-gram, 20 merkin ikkunointi).

Merkkivirhetason vaikutukset näkyvät osittain puhutuissa kysymyksissä. Lisäksi on muistettava että myös haettava materiaali on puheesta tunnistettu saman tunnistimen avulla. Siksi 120% merkkivirhetasollakin voidaan löytää relevantteja dokumentteja ja siksi toisaalta pienetkään merkkivirhetasot eivät takaa täydellistä hakutulosta.

Tekstimuotoisen hakusanan käyttö

Puheentunnistimen toteutuksessa on oletettu että suomen kielen kirjoitus- ja puheasu ovat yhteneviä. Tämän oletuksen seurauksena tunnistimen koulutuksessa on käytetty tekstimateriaalia. Myös yksittäisten sanojen merkkivirhetason laskenta perustuu tunnistimen tavoitteeseen tuottaa puheesta tekstiä. Edellisessä luvussa todettiin, että hakusanan merkkivirhetason pieneneminen yleensä parantaa suodatuksen tarkkuutta. Tulokset indikoisivat, että puhemateriaalin suodattaminen sanojen kirjoitusasun perusteella olisi mahdollista. Seuraavaksi selvitetään, miten kirjoitettujen hakusanojen käyttäminen vaikuttaa suodatukseen.

Kokeita varten puhuttujen hakusanojen sijasta käytettiin kirjoitettuja hakusanoja. Näistä muodostettiin nimikirjoitukset samalla tavalla kuin puhuttuja hakusanoja käytettäessä. Kuviossa 16 on esitetty suodatuksen keskimääräinen tarkkuus puhutuille (1-versio) sekä kirjoitetuille hakusanoille.



Kuvio 16. Keskimääräiset tarkkuudet puhutuille ja kirjoitetuille hakusanoille.

Kuviosta huomataan, että kirjoitettu hakusana parantaa suodatuksen keskimääräistä tarkkuutta kaikilla menetelmillä. Nyt huomataan myös että merkkivirhetason putoaminen (kirjoitetun hakusanan merkkivirhetaso on luonnollisesti 0%) vähentää parhaimpien menetelmien välisiä eroja: 4-grammien keskimääräiset tarkkuudet 20 ja

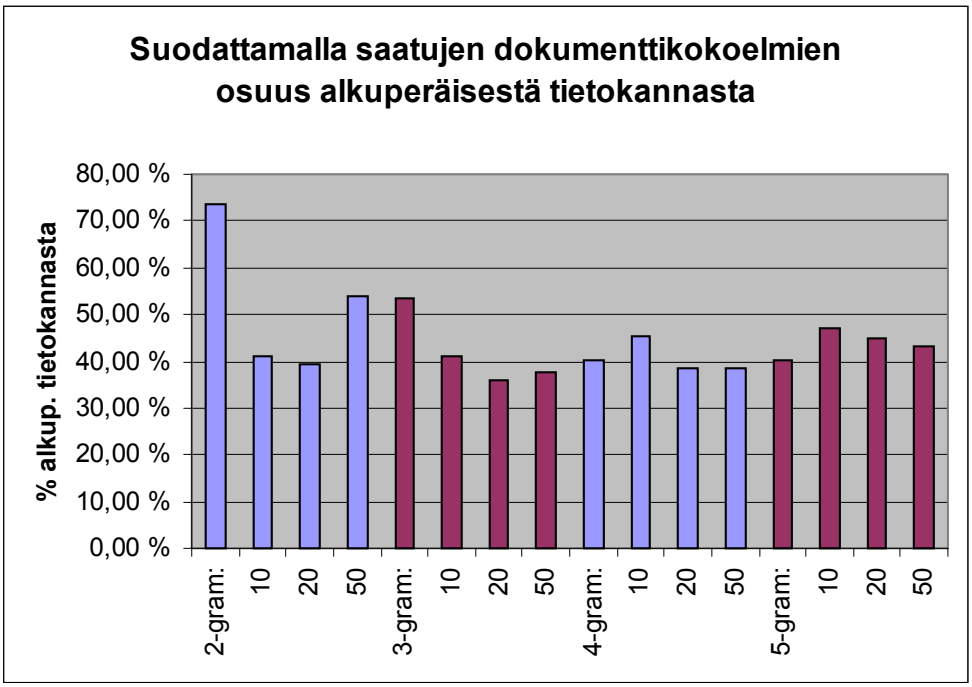
50 merkin ikkunalla eroavat parhaimmasta 3-grammista alle 0,5% keskimääräisen tarkkuuden verran. Tulokset viittaavat kuitenkin siihen suuntaan että n-grammeissa käytettyä *n* kokoa saattaa olla syytä kasvattaa kolmesta neljään, jos myös haettavien puhedokumenttien merkkivirhetaso saadaan laskettua.

6.5.5 Suodatuksen tehokkuudesta

Alaluvuissa 6.5.1–6.5.4 on vertailtu erilaisten n-grammimenetelmien soveltuvuutta puhemateriaalin suodatukseen sekä tarkasteltu suodatukseen vaikuttavia tekijöitä. Seuraavaksi arvioidaan, millaiseen suodatuskykyyn esitettyjen menetelmien avulla päästään. Tavoitteena on selvittää, millainen merkitys aiemmin esitetyillä menetelmillä olisi varsinaisen puhehaun järjestelmän toiminnan kannalta. Erityisen kiinnostavaa on selvittää, kuinka paljon suodatusmenetelmien avulla voidaan pienentää haun jatkokäsittelyssä mukana olevan dokumenttikokoelman kokoa.

Tutkimuksessa käytetty tietokanta sisältää 288 puhuttua uutista. Myös hakusanoista tiedetään, että tietokanta sisältää jokaista hakusanaa kohden keskimäärin 8,6 relevanttia dokumenttia. Näiden tietojen avulla voidaan arvioida millaiseen tarkkuuteen päästään, jos vastaus tuotetaan siten että kaikki tietokannan sisältämät dokumentit listataan satunnaisessa järjestyksessä hakutulokseen. *Tietokannan tarkkuus* on siis relevanttien dokumenttien keskimääräinen määrä tietokannassa jaettuna tietokannan koolla: $8{,}6/288 = 0{,}02986 \approx 3.0\%$.

Samalla laskentaperiaatteella voidaan myös laskea se hypoteettinen tietokannan osa, joka kattaa kaikki relevantit dokumentit sen jälkeen kun relevantit dokumentit ensin suodatetaan tuloslistaan kärkeen. Tämä luku saadaan selville jakamalla relevanttien dokumenttien keskimääräinen lukumäärä tietyn suodatusmenetelmän lopullisella tarkkuudella. Näin saadaan luku, joka ilmaisee sen dokumenttikokoelman (keskimääräisen) koon, johon sisältyy kaikki hakusanan kannalta relevantit dokumentit. Kuviossa 17 on esitetty, miten 2-, 3-, 4- ja 5-grammisuodatus kokonaisille uutisille sekä eri ikkunan koille vaikuttaa suodatuksen jälkeen käsiteltävään dokumenttikokoelman kokoon. Kuviossa on esitetty suodattamalla saatujen kokoelmien dokumenttien määrä prosenttiosuuksina alkuperäisen tietokannan koosta (288 dokumenttia).



Kuvio 17. Suodattamalla saatujen dokumenttikokoelmien prosentuaaliset osuudet alkuperäisen tietokannan koosta.

Kuviosta 17 huomataan, että parhaimmilla suodatusmenetelmillä poimittu dokumenttimäärä on alle 40% alkuperäisen tietokannan koosta. Parhaalla menetelmällä (3-gram, ikkunankoko 20 merkkiä) käsiteltävän tietokannan osa on suodatuksen jälkeen 35,8% alkuperäisestä dokumenttimäärästä.

Yllä esitetyssä vertailussa on oletettu, että mielenkiintoinen dokumenttijoukko kattaa ihannetilanteessa kaikki hakukysymyksen kannalta relevantit dokumentit. Kaikissa tilanteissa ei kuitenkaan ole ensisijalla tuottaa mahdollisimman kattava hakutulos, vaan käyttäjälle riittää yhden ainoan dokumentin löytäminen. Suodatusmenetelmän yksi etu on siinä, että päätös dokumenttikokoelman rajaamisesta voidaan tehdä jopa yksittäisen hakutilanteen mukaan. Toisin sanoen dokumenttikokoelma voidaan halutessa rajata eri saantitasojen mukaan eri kyselyissä. Tällöin lopullisen tarkkuuden sijasta voidaan esimerkiksi tarkastella tilannetta, jossa keskimäärin puolet tai neljäsosa tietokannan relevanteista dokumenteista on löydetty.

Menetelmien suodatuskykyä arvioidessa kannattaa kuitenkin pitää mielessä, että järjestelmän suorituskyky vaihtelee suuresti yksittäisten hakusanojen välillä. Edellä menetelmien vertailuun on käytetty keskiarvoja, joten annettuja prosenttiosuuksia ei soraan kannata soveltaa järjestelmän toteutuksessa. Yksittäisen suodatuksen

tehokkuuteen vaikuttaa tietokannan aineisto sekä hakusanat, joten näiden aiheuttamat vaihtelut suodatuskyvyssä pitäisi huomioida myös järjestelmän toteutuksessa.

Toisen syyn varovaisuuteen antaa koeasetelmassa käytetyn tietokannan keinotekoinen luonne. Koeasetelmaa varten pyrittiin luomaan mahdollisimman realistinen tietokanta, mutta tietokannan uutiset koskevat kaikesta huolimatta vain 17 eri aihetta. On mahdollista, että todellisella tietokannalla, jonka sisältämät dokumentit käsittelevät lukuisia eri aiheita, ei saavuteta yllä esitettyjä suodatusprosentteja. Sen sijaan menetelmien keskinäiseen paremmuusjärjestykseen dokumenttien aiheiden runsaus ei pitäisi vaikuttaa. (Siihen saattaa sen sijaan vaikuttaa puheentunnistukseen käytetty järjestelmä.)

6.6 Tulosten arviointi

Kansainvälinen puhehaun tutkimus on pääasiallisesti keskittynyt englanninkielisen puheen tutkimiseen. Useita vuosia kestäneen kehityksen jälkeen puhehaku on sellaisessa vaiheessa, että tutkimuksessa voidaan keskittyä mahdollisimman hyvien hakutulosten tuottamiseen. Suomenkielisestä puhehausta ei ole tehty aiempaa tutkimusta. Tämä tutkimus on siten ensiaskel suomenkielisen puhehaun tutkimuksessa. Rajattujen resurssien vuoksi tämän työn käytännön osuudessa on keskitytty hyvin pieneen puhehaun osa-alueeseen, nimittäin puhedokumentteja sisältävän tietokannan suodatukseen.

Suodatusmenetelmien avulla saavutettuja keskimääräisiä tarkkuuksia tarkastellessa on ilmeistä, että menetelmät eivät sellaisenaan sovellu puhetiedonhaun hakutulosten tuottamiseen. Järjestelmä nimittäin palauttaa moninkertaisen määrän epärelevantteja dokumentteja suhteessa relevantteihin dokumentteihin. Etenkin puhehaussa on erittäin tärkeä välttää epärelevantteja dokumentteja hakutuloksessa, koska dokumenttien kuunteleminen on hidasta. Erotuksena kansainvälisestä tutkimuksesta tuloksena syntynyt suodatettu tietokanta ei siis sellaisenaan ole vastaus hakijan esittämään hakupyyntöön, vaan työssä esitetyt menetelmät ovat väliaskel ennen lopullisen hakutuloksen tuottamista.

On myös paikallaan korostaa, että suodatuksessa ei tämän tutkimuksen yhteydessä ole käytetty kokonaisia hakupyyntöjä, vaan suodatus suoritettiin yksittäisten hakusanojen avulla. Yksittäisten hakujen keskimääräiset tarkkuudet vaihtelevat varsin paljon (ks. luku 6.5.4), samoin kuin käytettyjen hakusanojen tunnistustarkkuudet. On perusteltua olettaa, että useampaa hakusanaa käyttämällä

päästään parempiin tuloksiin. Ylipäätään se, että haluttuja dokumentteja kuvaa suurempi joukko sanoja, lisää järjestelmän suodatuksen aikana käytettävissä olevaa tietoa. Lisäksi on varsin epätodennäköistä että hakupyyntö sisältää pelkästään sellaisia sanoja, jotka kaikki tuottavat keskimääräistä heikompia hakutuloksia. Sanojen määrän lisääntyessä kasvaa lisäksi todennäköisyys, että ainakin osa hakusanoista on tunnistettu samalla tavalla kuin jokin sanan esiintymistä tietokannan dokumentissa.

Koska tutkimus eroaa kansainvälisestä tutkimuksesta sekä kohteena olevan kielen, käytettyjen menetelmien että koeasetelman tavoitteiden suhteen, käsillä olevassa tutkimuksessa esitettyjä arvoja ei ole mielekästä verrata kansainvälisen tutkimuksen tuottamiin tarkkuuslukuihin. Jotain yhteyksiä kansainväliseen tutkimukseen kuitenkin on vedettävissä: tulokset osoittavat, että myös suomenkielellä päästään parhaimpiin tuloksiin käyttämällä 3 ja 4 merkin mittaisia n-grammeja (vrt. luku 5.3.2). Toisaalta tulokset osoittavat poikkeavasti, että 5-grammit antavat parhaimman tuloksen kun käsitellään kokonaisista uutisista muodostettuja nimikirjoituksia. Tämän tutkimuksen valossa ei kuitenkaan ole mahdollista määritellä, johtuvatko tulokset tunnistimen luonteesta vaiko käytetystä suodatusmenetelmästä.

Tutkimus vahvistaa olettamuksen, jonka mukaan n-grammien avulla voi hakea myös taivutusmuotoisia hakusanojen esiintymiä. Toisin sanoen taivutusmuotoa voidaan puhehaussa käsitellä lisävirheenä ja luottaa siihen, että eri taivutusmuodoissa esiintyvien hakusanojen erot käsitellään osittaistäsmäytyksen avulla. Lisäksi tutkimus osoitti myös, että tekstimuotoisia hakusanoja voidaan käyttää sellaisenaan puhemateriaalin suodatuksessa. Tässä mielessä suomenkielisen materiaalin käsittely eroaa huomattavasti esimerkiksi englanninkielisestä puhehausta. Englantia käsittelevässä järjestelmässä tekstimuotoisen hakusanan käyttö äännetunnistettujen puhetiedostojen hakemiseen edellyttäisi väistämättä hakusanojen äänneasun selvittämisen ennen haun suorittamista. Piirre ei todennäköisesti juurikaan vaikuta hakupyynnön käsittelyn tehokkuuteen. Sen sijaan se vähentää suomenkielisen puhehakujärjestelmän rakentamiseen vaadittavien osien määrää, koska järjestelmään ei tarvitse sisällyttää sanojen ääntämysperiaatteita koskevaa säännöstöä tai sanakirjaa.

7 Yhteenveto ja johtopäätökset

Tässä tutkielmassa olen käsitellyt suomenkielistä puhehakua ja sen erityispiirteitä. Suomenkielisen puhutun materiaalin hakuun soveltuvaa hakujärjestelmää ei ole aiemmin toteutettu, eikä suomenkielisestä puhehausta ole olemassa tätä edeltävää kokeellista tutkimusta. Puhutun materiaalin käsittely on kielisidonnaista, joten kansainvälisen tutkimusyhteisön lähinnä englanninkielen käsittelyyn kehittämät järjestelmät eivät sellaisenaan taivu suomenkielisen puheen hakemiseen.

Kansainvälisessä puhehaun tutkimuksessa on ollut vallalla kaksi päälähestymistapaa (ks. luku 5). Toisessa lähestymistavassa puhedokumentit on ensin tunnistettu tekstiksi laajan sanavaraston puheentunnistimen avulla, jonka jälkeen transkriptioihin on sovellettu tekstitiedonhaun menetelmiä. Vaihtoehtoisesti puheesta on sanojen sijasta tunnistettu yksittäisiä äänteitä tai äännejonoja. Tällöin transkriptioiden käsittelyssä on käytetty osittaistäsmäyttäviä menetelmiä vertaamaan hakusanan äänneasua tunnistimen tuottamiin transkriptioihin. Englanninkieltä käsittelevien tutkimusten tulokset osoittavat, että parhaimpiin tuloksiin puhehaussa päästään kun järjestelmä on yhdistelmä näistä kahdesta menetelmästä. Yksinään käytettynä sanakirjapohjaista tunnistusta hyödyntävät järjestelmät suoriutuvat äänteiden osittaistäsmäytykseen perustuvia järjestelmiä paremmin.

Suomenkielisen puhemateriaalin käsittelyä rajoittaa se, että vielä ei ole olemassa laajaan sanakirjaan pohjautuvaa suomenkielistä puheentunnistinta. Tällaisen puheentunnistimen kehittäminen on hyvin suuritöinen projekti. Suomenkielellä on lisäksi sellaisia piirteitä, jotka tekevät sen käsittelyn englanninkielistä puhetta vaikeammaksi. Sanojen morfologinen variaatio kasvattaa sanakirjapohjaisen puheentunnistimen leksikkoa, jolloin puheentunnistaminen todennäköisesti vaikeutuu (ks. luku 3.4). Lisäksi sanojen morfologinen variaatio aiheuttaa merkkijonotasolla eroavuuksia aiheeltaan samanlaisten sanojen täsmäytyksessä (ks. luku 4.2). Suomenkielen fonologia puolestaan saattaa vaatia uusien menetelmien kehittämisen kaksoisvokaalien ja -konsonanttien tunnistamiselle (ks. luku 3.4).

On kuitenkin mahdollista, että suomenkielessä on myös sellaisia piirteitä, jotka helpottavat puhehaun tehtävää verrattuna englanninkielisen puheen hakemiseen. Tässä työssä tehdyn kokeellisen tutkimuksen perusteella vaikuttaisi siltä, että puhehakujärjestelmään rakennettujen yksinkertaisten osittaistäsmäytysmenetelmien

avulla voidaan hakea myös taivutusmuotoisia hakusanan esiintymiä. Lisäksi tutkimus osoittaa, että puhedokumentteja voidaan käsitellä tekstimuotoisilla hakusanoilla ilman, että hakusana ensin saatetaan äännemuotoonsa.

Työssä esitetty kokeellinen tutkimus on vasta ensiaskel kohti suomenkielistä puhehakujärjestelmää. Tutkimuksen tarpeisiin rakennettiin puhehaun testausympäristö, jossa verrattiin n-grammeihin perustuvien nimikirjoitusten soveltuvuutta puhemateriaalin suodattamiseen.

Suodatusmenetelmät eivät sellaisenaan tuota hakutulokseksi kelpaavia tuloksia. Sen sijaan suodatus tuo puhedokumenttien käsittelyyn ensimmäisen askeleen, jolla voidaan aloittaa hakutuloksen tuottaminen. Tällöin tietokannasta voidaan rajata pois joukko epärelevantteja dokumentteja ja näin rajoittaa tarkempien hakumenetelmien avulla läpikäytävää dokumenttijoukkoa. Suodatuksen jälkeen voidaan näin soveltaa esimerkiksi puheentunnistimen ominaisuuksia huomioon ottavia editointietäisyyteen perustuvia algoritmeja, jotka kokonaiseen tietokantaan käytettyinä olisivat epäkäytännöllisen hitaita.

Tutkimuksesta saadut tulokset tukevat sitä olettamusta, että tietokannan esikäsittely n-grammeihin perustuvalla suodatuksella on kannattavaa. Tulokset osoittavat, että suodatus voi pienentää puhehaussa jatkokäsittelyssä käytettyjen algoritmien läpikäymää dokumenttimäärää melkein kolmannekseen alkuperäisen tietokannan koosta. Tutkimus antaa myös vertailevaa dataa eri *n* arvoilla n-grammien avulla muodostettujen nimikirjoitusten suodatuskyvystä, kun nimikirjoituksen muodostetaan kokonaisista uutisista tai pienemmistä uutisten osista.

Alati kasvavien puhetietokantojen nopeiden käsittelymenetelmien kehittäminen tulee olemaan välttämätöntä niin kauan kuin sanakirjapohjaiseen puheentunnistukseen perustuvia indeksitiedostoja ei ole käytettävissä. Edelleenkin kun sanoja tunnistavat järjestelmät tulevat markkinoille, tarvitaan äännetunnistukseen perustuvia järjestelmiä leksikon ulkopuolisten sanojen etsimiseen, sillä nämä sanat saattavat olla hakujen kannalta erityisen tärkeitä Lisäksi äännetunnistuksen nopeus sanakirjapohjaiseen tunnistukseen verrattuna tekee siitä varteenotettavan tekniikan myös tulevaisuudessa.

8 Kirjallisuus

Alkula, R. 2000. Merkkijonoista suomen kielen sanoiksi: Suomen kielen morfologisten tulkintaohjelmien liittäminen tekstitiedonhakujärjestelmään ja liittämisen vaikutukset tekstin tallennukseen ja hakuun. Väitöskirja. Tampereen yliopisto.

Alkula, R. & Honkela, T. 1992. Tekstin tallennus- ja hakumenetelmien kehittäminen suomen kielen tulkintaohjelmien avulla. VTT Julkaisuja 765. Valtion Teknillinen Tutkimuskeskus.

Allan, J. 2002. Perspectives on Information Retrieval and Speech. Teoksessa Coden, A.; Brown, W. & Srinivasan, S. (toim.) Information Retrieval Techniques for speech Applications. Springer. 1–10.

Allan, J.; Callan, J.; Croft, W.; Ballesteros, L.; Byrd, D.; Swan, R. & Xu, J. 1998. INQUERY does Battle with TREC-6. Proc. 6th Text Retrieval Conference (TREC-6). NIST Special Publications 500-242.

Amir, A.; Efrat, Al & Srinivasan, S. 2001. Advances in Phonetic Word Spotting. Proc. 10th International Conference on Information and Knowledge Management. ACM Press. 580–582.

Ashford, J. & Willett, P. 1988. Text Retrieval and Document Databases. Chartwell-Bratt.

Barnett, J.; Anderson, S.; Broglio, J.; Singh, M.; Hudson, R. & Kuo, S. 1997. Experiments in spoken queries for document retrieval. Proc. Eurospeech. 1323–1326.

Brown, M.; Foote, J.; Jones, G. Spärck Jones, K. & Young, S. 1996. Open-Vocabulary Speech Indexing for Voice and Video Mail Retrieval. Proc. ACM Multimedia. 307–316.

Chen, A.; He, J.; Xu, L.; Gey, F. & Meggs, J. 1997. Chinese text retrieval without using a dictionary. Proc. 20th ACM SIGIR. ACM Press. 42 – 49.

Chen, B.; Wang, H. & Lee, L. 2001. Improved Spoken Document Retrieval by Exploring Extra Acoustic and Linguistic Cues. Proc. 7th Eurospeech. CD-ROM.

Conover, W. 1980. Practical Nonparametric Statistics. John Wiley & Sons. 2 laitos (1971).

Crestani, F. 2001. Towards the use of prosodic information for spoken document retrieval. Proc. 24th ACM SIGIR. ACM Press. 420 - 421

Dharanipragada, S.; Franz, M. & Roukos, S. 1998. AudioIndexing for Broadcast News. Proc. 6th Text Retrieval Conference (TREC-6). NIST Special Publications 500-242. 115–119.

Efthimiadis, E. 1996. Query Expansion. Annual Review of Information Science and Technology. Teoksessa Williams, M. (toim.) Annual Review of Information Systems and Technology (ARIST) 31. 121-187.

Ferrieux, A. & Peillon, S. 1999. Phoneme-level Indexing for Fast and Vocabulary-Independent Voice/Voice Retrieval. Proc. ESCA ETRW Workshop on Accessing Information in Spoken Audio. 60–63.

Fujii, H. & Croft, W. 1993. A comparison of indexing techniques for Japanese text retrieval. Proc. 16th ACM SIGIR. ACM Press. 237 – 246.

Garofolo, J.; Voorhees, E.; Auzanne, C. & Stanford, V. 1998. Spoken Document Retrieval: 1998 Evaluation and Investigation of New Metrics. Proc. ESCA ETRW workshop on Accessing Information in Spoken Audio. 1-7.

Garofolo, J.; Auzanne, C. & Voorhees, E. 2000. The TREC Spoken Document Retrieval Track: A Success Story. Proc. 8th Text Retvieval Conference (TREC-8). NIST special publication. 107–130.

Gauvain, J.; Lamel, L. & Adda, G. 2000. Transcribing Broadcast News for Audio and Video Indexing. Comm. ACM 43(2). 64–70.

Geutner, P.; Finke, M. Scheytt, P.; Waibel, A. & Wactlar, H. 1998. Transcribing Multilingual Broadcast News using Hypothesis Driven Lexical Adaptation. Proc. DARPA Broadcast News Transcription and Understanding Workshop. Saatavilla verkosta osoitteesta http://www.informedia.cs.cmu.edu/mli/papers/darpa-Bcast-ws98-petra.pdf. Saatavuus tarkistettu 1.4.2003.

Hall, P. & Dowling, G. 1980. Approximate String Matching. ACM Computing Surveys, 12 (4). 381–402.

Hansson, P. 2000. The Effect of Individual Words' Information Status on Accentuation. Proc. Nordic Prosody. 89-101.

Hull, D. 1993. Using Statistical Testing in the Evaluation of Retrieval Experiments. Proc 16th ACM SIGIR. 329–338.

Hyyrö, H. 2000. Merkkijonotäsmäyksestä. Pro Gradu -tutkielma. Tampereen yliopisto.

Iivonen, A. 2000. Intonation of Finnish Questions. Proc. Nordic Prosody. 137–151.

IPA 1999. Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge University Press.

Itkonen, E. 2001. Maailman kielten erilaisuus ja samuus. Yleisen kielitieteen julkaisuja, 4, Turun yliopisto. 2. laitos (1997).

James, D. 1995. The Application of Classical Information Retrieval Techniques to Spoken Documents. Väitöskirja, Cambridge University.

James, D. 1996. A System for Unrestricted Topic Retrieval from Radio News Broadcasts. Proc. ICASSP. 279–282.

Jones, G.; Foote, J.; Spärck Jones, K. & Young, S. 1996. Retrieving spoken documents by combining multiple index sources. Proc 19th ACM SIGIR. ACM Press. 30–38.

Jourlin, P.; Johnson, S.; Spärck Jones, K. & Woodland, P. 1999. General Query Expansion Techniques for Spoken Document Retrieval. Proc. ESCA ETRW Workshop on Accessing Information in Spoken Audio. 8-13.

Järvelin, K. 1995. Tekstitiedonhaku tietokannoista: Johdatus periaatteisiin ja menetelmiin. Suomen ATK-Kustannus.

Järvelin, K. 1993. Merkkijonot, sanat, termit ja käsitteet informaation haussa. Kirjastotiede ja informatiikka, 12(4). 119-128.

Karlsson, F. 1998. Yleinen kielitiede. Helsingin yliopistopaino. Uudistettu laitos (1994).

Kivimäki, J.; Lahti, T. & Koppinen, K. 2000. A Phonetic Vocoder for Finnish. Proc. 10th European Signal Processing Conference (EUSIPCO).

Ladefoged, P & Maddieson, I. 1996. The Sounds of the World's Languages. Blackwell Publishers.

Lahti, A. 2002. Tuotepäällikkö, Lingsoft Oy. Henkilökohtainen sähköpostitiedonanto 23.12.2002.

Lee, J. & Ahn, J. 1996. Using n-grams for Korean text retrieval. Proc. 19th ACM SIGIR. ACM Press. 216–224.

Leppävirta, J. 2001. Verkko-opetuksen uudet muodot. eOpetus-hankkeeen julkaisu. Espoo. Saatavilla verkosta osoitteesta: http://130.233.158.46/eopetus/opaskirja.pdf Saatavuus tarkistettu 1.4.2003.

Lyons, R. 1998. Understanding Digital Signal Processing. Addison-Wesley.

McGurk, H. & MacDonald, J. 1976. Hearing lips and seeing voices. Nature 264. 746-748.

Navarro, G. 2001. A guided tour to approximate string matching. ACM Computing Surveys, 33(1). ACM Press. 31 – 88.

Ng, C. & Zobel, J. 1998. Speech Retrieval using Phonemes with Error Correction. Proc. 21st ACM SIGIR. ACM Press. 365–366.

Ng, C.; Wilkinson, R. & Zobel, J. 2000. Experiments in Spoken Document Retrieval using Phoneme N-grams. Speech Communication, 32(1–2). 61–77.

Ng, K. & Zue, W. 1997. Subword Unit Representations for Spoken Document Retrieval. Proc. Eurospeech. 1607–1610.

Ng, K. 2000. Information Fusion for Spoken Document Retrieval. Proc. 25th International Conference on Acoustics Speech and Signal Processing (ICASSP).

Nie, J.; Gao, J; Zhang, J. & Zhou, M. 2000. On the use of words and n-grams for Chinese information retrieval. Proc. 5th International Workshop on Information retrieval with Asian languages. ACM Press. 141 – 148.

Nienstedt, W.; Hänninen, O.; Arstila, A. & Björkqvist, S. 1992. Ihmisen fysiologia ja anatomia. WSOY.

Nöth, E. 1991. Prosodische Information in der automatischen Spracherkennung: Berechnung und Anwendung. Max Niemeyer Verlag.

Padmanabhan, M. & Picheny, M. 2001. Current State of the Art in Large Vocabulary Automatic Speech Recognition Algorithms. IBM Research Report RC 22185 (W0109-057).

Pirkola, A. 2001. Morphological Typology of Languages for IR. Journal of Documentation, 57(3). 330–348.

Pusateri, E. & Thong, J. 2001. N-best List Generation using Word and Phoneme Recognition Fusion. Proc. Eurospeech.

Roach, P. 1983. English Phonetics and Phonology. Cambridge University Press.

Robertson, A. & Willett, P. 1992. Searching for historical word-forms in a database of 17th-century English text using spelling-correction methods. Proc 15th ACM SIGIR. ACM Press. 256 – 265.

Robertson, A. & Willett, P. 1998. Applications of N-grams in Textual Information Systems. Journal of Documentation, 54(1). 49–69.

Robinson, T.; Hochberg, M. & Renals S. 1994. IPA: Improved Phone Modelling with Recurrent Neural Networks. In Proc. 19th International Conference on Acoustics Speech and Signal Processing (ICASSP). 37-40.

Robinson, T.; Hochberg, M & Renals, S. 1996. The use of recurrent networks in continuous speech recognition. Teoksessa Lee, C.; Paliwal, K. & Soonkg, F. (toim.). Automatic Speech and Speaker Recognition - Advanced Topics. Kluwer Academic Publishers. 233–258.

Robinson, T. 1998. Speech Analysis. Luentomateriaali. Saatavilla verkosta osoitteesta: http://svr-www.eng.cam.ac.uk/~ajr/SA95/. Saatavuus tarkistettu 1.4.2003.

Salton, G. & McGill, M. 1983. Introduction to Modern Information Retrieval. McGraw-Hill.

Sanderson, M. & Crestani, F. 1998. Mixing and Merging for Spoken Document Retrieval. Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries. 397–407.

Schäuble, P. & Wechsler, M. 1995. First Experiences with a System for Content Based Retrieval of Information from Speech Recordings. Teoksessa M. Maybury (toim.). Working notes. Proc. International Joint Conference on Artificial Intelligence (IJCAI) Workshop. 59–69.

Shannon, C. 1948. A Mathematical Theory of Communication. The Bell System Technical Journal, 27. 379-423, 623-656.

Siegler, M. 1999. Integration of Continuous Speech Recognition and Information Retrieval for Mutually Optimal Performance. Väitöskirja, Carnegie Mellon University.

Singhal, A.; Choi, J.; Hindle, D.; Lewis, D. & Pereira, F. 1998. AT&T at TREC-7. Proc. TREC-7. NIST Special Publications 500-242. 239–252.

Singhal, A. & Pereira, F. 1999 Document expansion for speech retrieval. Proc. 22nd ACM SIGIR. ACM Press. 34–41.

Smeaton, A.; Morony, M. Quinn, G. & Scaife, R. 1998. Taiscéalaí: Information Retrieval from an Archive of Spoken Radio News. Proc. 2nd European Conference on Research and Advanced Technology for Digital Libraries. 397-407.

Sormunen, E. 2000. A Method for Measuring Wide Range Performance of Boolean Queries in Full-Text Databases. Väitöskirja. Tampereen yliopisto.

Spärck Jones, K. ; Jones, G.; Foote, J. & Young, S. 1996. Experiments in Spoken Document Retrieval. Information Processing & Management, 32(4). 399–417.

Srinivasan, S. & Petkovic, D. 2000. Phonetic Confusion Matrix Based Spoken Document Retrieval. Proc. 23rd ACM SIGIR. ACM Press. 81–87.

Tilastokeskus 2003a. Yksityisten radioasemien keskimääräiset viikkotunnit 1992 – 2001. Saatavilla verkosta osoitteesta http://www.tilastokeskus.fi/tk/el/uarad4002.xls . Saatavuus tarkistettu 1.4.2003.

Tilastokeskus 2003b. Yleisradion radiolähetysten ohjelma-aika 1992 – 2001. Saatavilla verkosta osoitteesta http://www.tilastokeskus.fi/tk/el/uarad4000.xls . Saatavuus tarkistettu 1.4.2003.

Viikki, O. 1999. Adaptive Methods for Robust Speech Recognition. Väitöskirja. Tampereen teknillinen korkeakoulu.

Wechsler, M. 1998. Spoken Document Retrieval Based on Phoneme Recognition. Väitöskirja. Swiss Federal Institute of Technology (ETH), Zurich.

Wechsler, M & Schäuble, P. 1995 Speech Retrieval Based on Automatic Indexing. Final Workshop on Multimedia Information Retrieval (MIRO'95).

Weinschenk, S. & Barker, D. 2000. Designing Effective Speech Interfaces. John Wiley & Sons.

Whittaker, E.; Thong, J. & Moreno, P. 2001. Vocabulary Independent Speech Recognition Using Particles. Proc. Automatic Speech Recognition and Understanding Workshop (ASRU).

Wiik, K. 1981. Fonetiikan perusteet. WSOY.

Wiik, K. 1965. Finnish and English Vowels. Väitöskirja. Turun yliopisto.

Witbrock, M. & Hauptmann, A. 1997a Speech Recognition and Information Retrieval: Experiments in Retrieving Spoken Documents. Proc. DARPA Speech Recognition Workshop.

Witbrock, M. & Hauptmann, A. 1997b. Using Words and Phonetic Strings for Efficient Information Retrieval from Imperfectly Transcribed Spoken Documents. Proc. 2nd ACM International Conference on Digital Libraries. 30-35.