KIMMO KOSKENNIEMI

JOHDATUS KIELITEKNOLOGIAAN, SEN MERKITYKSEEN JA SOVELLUKSIIN

NYKYKIELTEN LAITOKSEN OPPIMATERIAALIA 1
Helsingin yliopiston nykykielten laitos 2013

NYKYKIELTEN LAITOKSEN OPPIMATERIAALIA 1

Kimmo Koskenniemi

Johdatus kieliteknologiaan, sen merkitykseen ja sovelluksiin

Helsingin yliopiston nykykielten laitos
PL 24 (Unioninkatu 40 B)
FI-00014 Helsingin yliopisto

www.helsinki.fi/nykykielet

ISBN 978-952-10-8677-9 (Pdf)
ISSN 2323-8828 (Verkkojulkaisu)

hdl.handle.net/10138/38336

Alkusanat

Tämä oppikirja on syntynyt yli kymmenen vuoden ajan pitämieni kieliteknologian johdantokurssien ohella. Kursseilla käytettyä versiota on joka kierroksella paranneltu ja eläkkeelle jäätyäni julkaisen sen vapaana julkaisuna [1] käytettäväksi ja samoin ehdoin edelleen kehitettäväksi jatkossa. Oppikirjaa on käytetty viikko kerrallaan etenevän yhden periodin mittaisen verkkokurssin materiaalina. Kurssi on pidetty yleensä kokonaan verkon kautta ilman kateederiluentoja tai muuta suoraa kontaktia luennoijan ja opiskelijoiden välillä. Kieliteknologian opetuksen valtakunnallisen verkoston aktiivisena kautena suorittajista suuri osa oli muista kuin Helsingin yliopistosta ja edelleenkin osallistujia on eri kaupungeista, satunnaisesti myös vaihdon takia ulkomailla olevia opiskelijoita.

Kurssin suorittamiseksi on ollut oheismateriaalina luettavia artikkeleita tai pätkiä kirjoista sekä viikkokysymyksiä, joilla on varmistettu se, että opiskelijat joka viikko omistavat riittävästi aikaa teemalle. Viikkokysymyksien lisäksi on ollut verkkoselaimen avulla suoritettava yksinkertaisten harjoitustehtävien paketti, jolla opiskelijat ovat käytännössä kokeilleet erilaisia kieliteknologisia sovelluksia tai hankkineet niistä muuten tietoja. Edistynyt opiskelija on toiminut kurssiassistenttina ja on tarkistanut viikkotehtävät että harjoitustyöpaketin. Kurssin suorittamiseksi on kuitenkin järjestetty perinteinen kirjallinen tentti, jonka perusteella viikkotehtävät ja harjoitustyöpaketin hyväk-

[1] *Creative Commons: Nimeä—Jaa Samoin* -lisenssin mukaisena ks. http://creativecommons.org/licenses/by-sa/3.0/deed.fi

syttävästi suorittaneet ovat saaneet suoritusmerkinnän ja arvostelun. Menettely on osoittautunut varsin toimivaksi sikäli, että useat sadat opiskelijat ovat kurssin suorittaneet. Opiskelijapalaute puolestaan on vahvistanut käsitystä, että työmäärä ja opintopisteet ovat olleet kohtuullisessa suhteessa toisiinsa.

Kirja on muotoiltu tietokoneen ruudulta helposti luettavaan muotoon siinä toivossa, että sitä ei tarvitsisi tulostaa paperille. Linkit verkkosivuille, asiahakemisto ja sisällysluettelo toimivat digitaalisessa versiossa aktiivisina linkkeinä.

Sisältö

Luku 1

Yleistä taustaa

Tämä oppikirja on laadittu erityisesti kieliteknologian sovellusten näkökulmasta ja se jäsentyy kuudeksi teemaksi, kukin oman sovellusalueensa ympärille. Kieliteknologian menetelmiä ja käsitteitä tuodaan esille sovellusten esittelyn yhteydessä. Tavoitteena on johdattaa lukija ymmärtämään kielen olemusta kieliteknologian kannalta sekä auttaa tunnistamaan kieliteknologian merkitys ja mahdollisuudet erilaisissa nykyään jo yleisissä, mutta tulevaisuudessa vieläkin yleisemmissä tärkeissä sovelluksissa.

1.1 Mitä kieliteknologia on ja mihin sitä tarvitaan?

Kieliteknologia on kieleen liittyvää teknologiaa ja se käsittelee erityisesti sellaisia menetelmiä, joilla *ihmisten kieltä voidaan tietokoneen avulla automaattisesti jäsentää, tuottaa tai tunnistaa* ja erityisesti myös menetelmiä, joiden avulla ihminen voi kielensä avulla kommunikoida tieto-

koneen kanssa. Ihmisten kieli voi olla joko kirjoitettua, puhuttua tai viitottua kieltä.

Kieliteknologian ydin koostuu nimenomaan niistä menetelmistä, joilla kieltä ja sen käyttöä mallinnetaan. Ytimen ympärillä on laaja joukko sovelluksia, joissa ytimeen liittyviä menetelmiä voidaan hyödyntää ja jotka tuovat kieliteknologian piiriin uusia tarkastelukulmia ja haasteita kieliteknologian menetelmin ratkaistaviksi tehtäviksi. Usein tätä laajempaa kokonaisuutta kutsutaan nimenomaisesti kieliteknologiaksi (ruotsiksi språkteknologi, englanniksi language technology, language engineering, human language technology) ja ydintä *tietokonelingvistiikaksi* (ruotsiksi datalingvistik tai datorlingvistik, englanniksi computational linguistics). Suomessa kieliteknologiatermiä käytetään nykyään molemmista.

Ajatus puhuvasta ja puhetta ymmärtävästä tietokoneesta on aika vanha. Kubrikin vuonna 1968 valmistuneessa elokuvassa *Avaruusseikkailu 2001* oli HALiksi kutsuttu puhuva (teko)älykäs tietokone ja vuodesta 1966 alkaen tehdyissä suosituissa *Star Trek* -filmeissä tietokonetta käytettiin luontevasti puhekäyttöliittymän kautta. Vuosikymmenien kuluessa tietokoneet ovat kehittyneet ja nopeutuneet valtavasti. Silti näin laaja ja luonteva käyttöliittymä on pysyy yhä tulevaisuuden toiveena. Tosin nykyään pystytään jo toteuttamaan monia sovelluksia, joissa ihminen voi kysellä ja toimittaa erinäisiä asioita tietokoneen kanssa. Kieli- ja puheteknologian tehtävänä on opettaa tietokone tunnistamaan ihmisen puhetta sekä tuottamaan ymmärrettävää ja riittävän luontevaa puhetta. Keskusteltaessa tietokoneen kanssa täytyy tietokoneelle näiden lisäksi opettaa myös miten keskustellaan eli mitä ja miten laajasti

milloinkin tulisi kysyä, todeta tai vastata.

Eurooppa on monikielinen ja se on poliittisestikin sitoutunut monikielisyyteen. Euroopan unionin melkoisessa byrokratiassa kielen *kääntäminen* muodostanee isoimman yksittäisen kuluerän ja unioni työllistää melkoisen joukon kääntäjiä ja tulkkeja. EU:lla lienee pitkälti toistatuhatta vakinaista kääntäjää ja lisäksi freelancereita. Vuosittain siellä käännetään toista miljoonaa sivua tekstiä EU-kielille tai -kieliltä. Tekstejä käännetään osittain ihmisvoimin perinteiseen tapaan, mutta myös sekä kieliteknologisten ja muiden menetelmien avustamana ja osittain myös automaattisten käännösohjelmien avulla. Tietokoneohjelman suorittama käännös yleensä tarkistetaan ja korjaillaan kuitenkin ihmisen toimesta. Kielen kääntämisessä kieliteknologiasta on suuri apu ihmiselle, joka työkseen tai työssään kääntää tekstejä.

Eri tavoin syntyvä tieto tallennetaan nykyään useimmiten valmiiksi suoraan tietokoneen muotoon. Uskomattoman paljon tietoa sijaitsee Internetissä kenen tahansa saatavilla tai erilaisissa yrityksissä ja organisaatioissa valikoidumpien käyttäjien nähtävissä. Tietoa on oikeastaan liikaa. Siksi tiedon valikointi eli *tiedonhaku*, *tiedon tiivistäminen, luokittelu* ja *indeksointi* ovat nousseet tärkeiksi. Kun tieto on kielen muodossa, tarvitaan kieliteknologiaa neutraloimaan esim. sanojen taipumisesta johtuvia ongelmia, jotta mahdollisimman monet halutuista dokumenteista löytyisivät. Toisaalta kielellisen rakenteen tunnistaminen auttaa monissa tapauksissa tarkentamaan käsittelyä eli tuomaan haussa enimmäkseen sellaisia dokumentteja, joita varsinaisesti halutaan.

Muidenkin kielten kuin äidinkielen oppiminen kuuluu

olennaisena osana eurooppalaiseen yhteiskuntaan. Tietokoneen ja kieliteknologian avulla voisimme helpottaa vieraskielisten tekstien ymmärtämistä ja avustaa ihmisiä vieraan kielen oppimisessa. Tietokoneet kun ovat paremmin saatavilla ajasta ja paikasta riippumatta kuin ihmisopettajat. Monien mielestä tietokone on myös hienotunteinen opettajana, kun sille tehtyä virhettä ei tarvitse häpeillä.

Myös vammaisten elämänlaatua voidaan (ja voitaisiin) monin tavoin parantaa kieliteknologian avulla, esimerkiksi puhesynteesi auttaa näkövammaisia tekstien lukemisessa ja erityiset ohjelmat voivat auttaa kommunikoimaan, vaikka puheentuottaminen ja kirjoittaminen olisi muuten vaikeaa.

Viime vuosituhannella kieliteknologia (tai tietokonelingvistiikka) oli lähinnä tutkimuskohde, tosin varsin mielenkiintoinen kohde. Viime aikoina tietoverkkojen ja tietotekniikan läpimurron myötä ollaan konkreettisestikin siirtymässä tietoyhteiskuntaan, jossa uudet viestintävälineet ja erityisesti Internet koskettavat jo pääosaa kaikista kansalaisista ja niiden avulla on saatettu ennennäkemättömät tietomäärät ja verkkopalvelut jokaisen ulottuville. Suuri osa näistä tiedoista on luonnollisen kielen muodossa eikä esimerkiksi numeroina tai kuvina. Moniin verkossa oleviin palveluihin olisi puhuttu (tai kirjoitettu) kieli luontevin lähestymiskeino. Mitenkä voisi pienen kännykän kautta luontevammin pyyntöjänsä esittää kuin puhumalla ja millaisessa muodossa kuin puheena vastauksia olisi pienellä laitteella mukavinta saada. Harvojen harrastuksesta on siten hyvin lyhyessä ajassa tullut massojen päivittäistä elämää ja toimintaa koskettava teknologiaa, jonka tulevaa merkitystä vieläkin ilmeisesti aliarvioidaan.

Lyhyt katsaus kieliteknologian olemukseen, tehtäviin ja historiaan löytyy laajasta englanninkielisestä oppikirjasta (Jurafsky & Martin, 2008, Introduction, pp. 9–16). Suomen- ja englanninkielinen katsaus kieliteknologiaan ja sen merkitykseen suomen kielen kannalta löytyy META-NET -hankkeen tuottamasta julkaisusta (Koskenniemi et al, 2012), joka on luettavissa vapaasti verkostakin.

1.2 Kielen järjestelmä

Ihmisen kieltä kuten suomea tai englantia voidaan kuvata järjestelmänä, mutta niiden järjestelmä on olemukseltaan jotain muuta kuin esimerkiksi tietokoneiden ohjelmointikieli tai kokoelma matemaattisia kaavoja. Ihmiskieli on sekä *laaja*, että *monimuotoinen*, mutta myös aivan erityisellä tavalla *sumea* ja rajoiltaan *avoin järjestelmä*. Ihmiskieliin verrattuna ohjelmointikielet ovat hyvin yksinkertaisia.

1.2.1 Kieltä ei tiedosteta

Ihmisellä ei luonnostaan ole intuitiivista kuvaa kielensä olemuksesta, ei sen monimutkaisuudesta, eikä monitulkintaisuudesta. Päinvastoin, useimmille kieli tuntuu jokseenkin itsestään selvältä ja ongelmattomalta. Äidinkielen kohdalla tällainen sokea pilkku on ehkä konkreettisempi kuin koululaisena tai aikuisena opiskeltujen vieraiden kielten. Useat kokevat äidinkielen paljon säännöllisemmäksi kuin, mitä se on. Vieraiden kielten tietyt poikkeuksellisuudet muistetaan helpommin. Silti, minkään kielen kohdalla sen oppiminen ei merkitse sitä, että henkilö itse pystyisi osaamisensa perusteella selittämään kielen sääntöjä tai

säännönmukaisuuksia kovinkaan tarkasti.

Osatakseen jotakin, esimerkiksi ajaa polkupyörällä tai kävellä, ihmisen ei tarvitse tietää, kuinka hän sen osaa. Taitoja vain opitaan, eikä ilmiöiden tarkempaa rakennetta tarvitsekaan tietää. Pyöräilijälle riittää, että pysyy pystyssä ja pääsee, minne haluaa. Ei ole siis pakko olla selvillä siitä, että kääntyäkseen oikealle pitää ensin saada pyörä kallelleen, esimerkiksi kääntämällä ensin hiukan vasemmalle, jonka jälkeen oikealle kääntyminen hoituu sillä, että pitää pyörän pystyssä ohjaamalla vaistonvaraisesti. Mutta, jotakin tällaista me teemme, vaikka emme tiedostaisikaan fysikaalisia tosiasioita, ja silti ajamme pyörää sujuvasti. Kävelemiseen tarvittavan mekaniikan ohjeistus olisi kai vielä mutkikkaampaa, mutta onneksi voimme kävellä kahdella jalalla ilman tällaista tietoa.

Kieleenkin liittyy tällaisia osaamisia, joita useimmat meistä eivät tiedosta. Meidän olisi vaikeaa ja hyvin työlästä kirjoittaa täsmällistä kielioppia äidinkielestämme, siitä millaiset sanajärjestykset ja sanavalinnat milloinkin ovat luontevia tai miten sanoja taivutetaan.

Jos ja kun kielen toimintaa yritetään ohjelmoida tietokoneelle tai muuten kuvata eksplisiittisesti, käy pian ilmeiseksi, että kieli on kovin laaja, moniselitteinen ja vaikeasti kuvattava kohde.

1.2.2 Lekseemi eli hakusana, sananmuoto ja sane

Tavallisessa kielenkäytössä *sana* voi tarkoittaa useammanlaisia asioita. Kieliteknologiassa ja kielitieteissä on tarpeen kuitenkin pitää tarvittaessa erillään seuraavat kolme

käsitettä, jotka ovat tämän kurssin kannalta tärkeitä ja sen vuoksi ne tuodaan esille jo nyt.

lekseemi eli hakusana: Intuitiivista sanan käsitettä lähinnä lienee *lekseemi* (jota usein kutsutaan sanakirjojen yhteydessä myös *hakusanaksi*). Sellaisilla voi olla taivutusmuotoja ja merkityksiä ym. Esim. KATTO-lekseemillä on taivutusmuotoja siten kuin millä tahansa substantiivilla, esim. KATTO, KATON, KATTOA, jne. Yksikielisissä sanakirjoissa kuten *Suomen kielen perussanakirjassa* hakusanalla on oma selitystekstinsä, jossa sen mahdollisia erilaisia alamerkityksiä kuvataan. Kaksikieliset sanakirjat pyrkivät puolestaan luettelemaan kielen hakusanat ja antamaan kullekin yhden tai useampia käännöksiä jollekin toiselle kielelle. Lekseemi on siis yleensä tällaista ns. hakusana-artikkelia vastaava kielen kuvaamisen yksikkö, johon liittyy perusmuodon lisäksi esim. tieto sanaluokasta ja taipumisesta. Siten esimerkiksi LAKI merkityksessä 'säädös' on eri lekseemi kuin LAKI merkityksessä 'laakea huippu' koska nämä kaksi taipuvat eri tavalla (LAIN vs. LAEN). Eri alamerkitys ei kuitenkaan tee eri lekseemejä eli esim. LASKEA erilaisissa merkityksissään on yksi ja sama lekseemi (ja sen alamerkitykset luetellaan samassa hakusana-artikkelissa). Lekseemit voivat olla yksiosaisia tai yhdyssanoja, kuten *harjakatto*, ne voivat myös olla johdettuja sanoja, kuten *ojentautua*.

sananmuoto: Lekseemin taivutusmuodot ovat *sananmuotoja*, esim. sananmuoto KATOLLE on KATTO-lekseemin yksikön allatiivi. Samaan tapaan kuin leksee-

mikin, sananmuoto on käyttöyhteydestään irrotettavissa oleva käsite. Siten KATTO-lekseemillä on vain yksi yksikön allatiivimuoto KATOLLE, vaikka tekstissä tuo kirjainjono esiintyisi useampia kertoja. Sananmuoto on vain merkkijono, esim. KATOSTA, jolla voi olla useampikin tulkinta, tässä se voi olla joko KATTO-lekseemin elatiivi tai yhtä hyvin KATOS-lekseemin partitiivi. Erityisesti perusmuodotkin ovat sanamuotoja, siis KATTO on myös sananmuoto, nimittäin KATTO-lekseemin yksikön nominatiivi.

sane: Juokseva teksti koostuu puolestaan *saneista*, joita erottaa toisistaan sananväli tai välimerkit. Voimme siten laskea kuinka monta sanetta jossakin tekstissä on. Sane on siten sananmuodon esiintymä, eli tietty sananmuoto, esim. ON voi esiintyä tekstissä vaikkapa 12 kertaa. Varsinkin tekstin pituudesta käytetty ilmaus ”teksti on 1000 juoksevan sanan mittainen” tarkoittaa meidän termeillämme ”1000 saneen mittainen”.

Vaikka olemme määritelleet nämä kolme käsitettä ja termiä, saatamme silti käyttää sujuvuuden vuoksi silloin tällöin termiä ’sana’ kussakin näistä merkityksistä sikäli, kun sekaantumisen vaaraa ei ole.

Lauseet ja virkkeet koostuvat saneista, joiden välillä on erilaisia ns. määritesuhteita. Virkkeiden rakenteen tunnistaminen voisi muuten olla kohtuullisen helppoa, mutta tätä tehtävää vaikeuttaa sananmuotojen ja rakenteiden moniselitteisyys. Virkkeiden merkitykset ovat niiden rakennettakin ongelmallisempia kuvata täsmällisesti, vaikka

ihmisillä onkin se vaikutelma, että merkitykset olisivat itsestään selviä.

1.2.3 Kieli on iso

Kieli ei ole pieni eikä yksinkertainen, vaikka sen koosta ja mutkikkuudesta meillä ei olekaan luontaista mielikuvaa. Päinvastoin, kieli on monella tavalla laaja tai ehkä ääretönkin. Tiedämme matematiikasta, että kokonaislukuja on äärettömän paljon, vaikka kukin luku koostuu jonosta numeromerkkejä, joita kymmenjärjestelmässämme on kymmenen erilaista: 0–9. Kielikin koostuu vastaavalla tavalla harvoista merkeistä: kirjoitettu kieli kirjaimista ja puhuttu kieli äänteistä. Jätämme tässä kutenkin hetkeksi puheen ja mutkikkaammat kirjoitusjärjestelmät syrjään ja käsittelemme suomen tapaista aakkosmerkeillä kirjoitettua kieltä.

Kirjaimia tai äänteitä on kielissä muutama kymmenen erilaista ja niiden voidaan ajatella vastaavan numeromerkkejä. Kirjaimilla voidaan muodostaa sananmuotoja, joskaan kaikki kirjainyhdistelmät eivät ole mahdollisia: TALOSSA ja SPRIIN ovat mahdollisia, mutta KDPGV tuntuu mahdottomalta suomen kielessä. Kaikki mahdollisen tuntuisetkaan sananmuodot kuten HEULO[1] eivät kuulu kieleen. Sananmuodoilla on myös rajallinen pituus, eikä minkään kielen sanavarasto ole loputon. Kielissä voi olla muutamia kymmeniä tuhansia tai jopa miljoona lekseemiä, jotka voivat olla lueteltuina hakusanoina sanakirjoissa. Vaikka lekseemien määrä on tässä mielessä rajallinen, elävien kielten sanasto on avoinna sekä pysyvämmin kie-

[1] Tämänniminen paikka lienee Havaijilla.

leen pyrkiville uudissanoille että puhujien muodostamille tarpeeseen luoduille lekseemeille, jotka saattavat unohtua käytön jälkeen.

Lekseemit eivät kuitenkaan ole kiinteitä yksiköitä, vaan useimmissa kielissä ne taipuvat sananmuotoina, minkä lisäksi monista lekseemeistä voi johtaa säännönmukaisesti toisia lekseemejä ns. *johdoksia*. Lekseemejä voidaan vielä yhdistää *yhdyssanoiksi*, jotka ovat nekin lekseemejä. Tällainen lekseemien tuottaminen toisista lekseemeistä eli sananmuodostus saattaa olla vain pieni lisä kielen sanavarastoon kuten englannin kielessä, tai sitten ratkaiseva tekijä kuten suomessa.

Suomen kielessä erityisesti sanojen taipuminen on yllättävänkin monimuotoista. Jokainen substantiivi saa erilaisia muotoja

- kahdessa luvussa (eli yksikössä ja monikossa),
- yli kymmenessä sijamuodossa (nominatiivi, genetiivi, partitiivi, jne.),
- omistusliitteen mukaan (yksikössä ja monikossa kolme persoonaa ja ilman liitettä) sekä
- liitepartikkelin mukaan (-kin, -pa, -han, jne).

Yhteensä näiden yhdistelmät tuottavat noin 2000 erilaista sananmuotoa kustakin substantiivista. Adjektiivit taipuvat samaan tapaan kussakin kolmesta eri vertailuasteestaan (jotka ovat positiivi, komparatiivi, superlatiivi). Näin ollen kullekin adjektiiville tulee noin 6000 eri muotoa. Verbit yltävät vieläkin useampiin muotoihin, peräti noin 12 000–18 000 muotoon, joista pääosa tulee partisiipeista ja muis-

ta nominaalimuodoista, jotka taipuvat kuten substantiivit (juokse+minen) tai joillakin verbeillä kuten adjektiivit (katso+ttu).

Minkälaisiin suuruusluokkiin tämä johtaa? Jos oletamme suomen kielessä olevan esim. 100 000 yksiosaista substantiivia, saamme näistä taivuttamalla 200 miljoonaa eri muotoa. Adjektiiveja on vähemmän, mutta niistä voinee tulla sata miljoonaa sananmuotoa. Verbejä on myös vähemmän, ja niistä voisi tulla vielä sata miljoonaa lisää.

Tässä lähes puolessa miljardissa sananmuodossa ei kuitenkaan ole koko totuus. Ensinnäkin voimme suomen kielessä johtaa verbeistä, adjektiiveista ja substantiiveista toisia hakusanoja, esim.: ISTUA, ISTUSKELLA, ISTUSKELUTTAA, ISTUTTAA, ISTUUTUA, ISTUSKELUTTAJAMAISUUS, jne. Tätä kautta saamme muotojen määrän ehkä yhtä kertalukua (eli kerrointa 10) suuremmaksi.

Isompi vaikutus on kuitenkin yhdyssanojen muodostamisella. Kahdesta substantiivista voi muodostaa yhdyssanan, esim. TALO ja KIRJA yhdistyy sanaksi TALOKIRJA tai TALONKIRJA. Kaksiosaisia yhdyssanoja voisi siis olla noin $2 \times 100\,000^2$ ja niillä kullakin ne 2000 muotoa, eli yhteensä 40 biljoonaa (siis 40×10^{12}). Yhdyssanojen muodostaminen ei kuitenkaan rajoitu kaksiosaisiin, vaan esimerkiksi ruokaloissa näemme useinkin sellaisia yhdyssanoja kuten JAUHELIHAMAKARONILAATIKKO, SAVUKIRJOLOHISALAATTI. Neliosaisten yhdyssanojen muotojen teoreettinen määrä kohoaakin jo kohtuuttoman suureksi: $100\,000^4 \times 2000$ eli

$$200\,000\,000\,000\,000\,000\,000\,000$$

Kun huomaamme, että yhdyssanan alkuosa voi olla jo-

	kerroin	*yhteensä erilaisia muotoja*
perusmuoto: KATTO		1
yksikkö ja monikko: KATTO, KATOT	2	2
sijamuodot: N, A, NA, KSI, SSA, STA, VN, LLA, LTA, LLE, TTA, INE, IN	13	26
omistusliitteet: NI, SI, VN, NSA, MME, NNE	7	182
liitepartikkelit: KIN, HAN, PA, KO, …	11	2 002
yksiosaisille substantiiveille: AALTO, AAMU, …, KATTO, …	90 000	180 180 000
kaksiosaisille yhdyssanoille: AALTO-PELTI, …	180 000	32 432 400 000 000
kolmiosaisille yhdyssanoille: AALTO-PELTI-KATTO, …	180 000	5 837 832 000 000 miljoonaa
neliosaisille yhdyssanoille: JAUHE-LIHA-MAKARONI-LAATIKKO	180 000	1 050 809 760 000 miljoonaa miljoonaa

Kuva 1.1: Suomen kielen sananmuotojen määrien suuruusluokkia

ko nominatiivissa tai genetiivissä, saamme tästä helposti vaikka lukumääriä 10^{24} eli kvadriljoonan, vertaa taulukko

1.1. Luvut ovat hyvin keinotekoisia kahdellakin tavalla. Toisaalta juuri neljä yhdyssanan osaa on mielivaltainen, joskus voidaan tehdä pitempiäkin. Tärkeää on huomata, että monet näistä muodollisesti mahdollisista kombinaatioista ovat vailla sovittua merkitystä tai käyttöä, esim. ÄÄNIKALAUNENHEIKKOUS tai JÄÄOVIPIIRAKANNAULA.

Kuitenkin jokainen suomen kielen taitaja eräässä mielessä hallitsee joka ikisen noista kvadriljoonasta sananmuodosta. Hän pystyy vaivattomasti tunnistamaan sellaisen osat ja taivutusmuodot, eli pystyy oitis todentamaan, onko muoto mahdollinen eli muodollisesti korrekti.

Sen lisäksi, että näitä suuria lukuja voi kummastella, näistä laskelmista voidaan tehdä eräs johtopäätös ihmisen kielikyvystä. Ei ole uskottavaa, että kielenpuhujat oppisivat valmiita sananmuotoja siten, että heidän tulee kuulla opittavana oleva sananmuoto ennen, kuin se tulee opituksi. Sananmuotoja on nimittäin liikaa. Kvadriljoonan sananmuodon luettelemiseen sananmuoto per sekunti tarvittaisiin enemmän sekunteja kuin, mitä maapallo on ollut olemassa. Maapallon iäksi kun arvioidaan noin 4 miljardia vuotta eli likipitäen $4 \times 10^9 \times 365 \times 24 \times 3600s$ eli noin $1,2 \times 10^{17}$ sekuntia.

Pieni muistutus on tässä paikallaan. Esimerkki sananmuotojen runsaudesta oli suomen kielestä. Meillä on usein houkutus kuvitella, että oma kielemme olisi jollakin tavoin äärimmäinen. Maailman muutaman tuhannen kielen joukossa se on kuitenkin monella tavalla keskiverto. Toiset kielet ovat sananmuodostukseltaan suomea yksinkertaisempia, jotkut taas monimutkaisempia. Esimerkiksi eskimokielissä yksi sananmuoto vastaa rakenteeltaan lä-

hestulkoon eurooppalaisten kielten lausetta. Sen vuoksi eskimokielissä voisi olla kertalukuja enemmän erilaisia sananmuotoja kuin suomessa. Sekä sanskritissa että klassillisessa arabiassa ovat sananmuotojen rakenteet ja taipumisen tai sananjohdon yhteydessä tapahtuvat vaihtelut paljon mutkikkaampia kuin suomen kielessä. Maailmassa puhutuista kielistä ks. Summer Institute of Linguistics -järjestön tuottamaa kirjaa (Lewis, 2009) tai sen vapaasti verkossa selattavaa versiota.

1.2.4 Kielikyky

Kielenpuhujalla sanotaan olevan *kielikyky* eli kompetenssi, jonka turvin hän kieltä ymmärtää ja käyttää. Kielikykyyn liitetään erityisesti myös *produktiivisuus* eli kyky ymmärtää ja tuottaa ilmauksia säännönmukaisuuksien perusteella ilman, että niitä on nimenomaisesti ennen kuultu, nähty tai opittu.

Kieli ei ole tarkkarajainen kohde, vaan *uusia sanoja* opitaan ja tehdään tarpeen mukaan ja kielen sääntöjä joskus venytetään suorastaan *leikiksi* asti. Arkielämässäkin kohtaamme uusia nimiä, joita osaamme taivuttaa yhteisten sopimusten mukaisesti. Näin, vaikka emme tiedostaisi noita yhteisiä sopimuksia, emmekä osaisi niitä pukea miksikään säännöiksi, kuten aiemmin todettiin.

Osaamme *myös taivuttaa sujuvasti myös uusia*, ennen kuulemattomia *sanoja*, nimiä tai uudissanoja. Esimerkiksi monikon genetiivin muodostaminen tekosanasta HEUHU on ilman muuta HEUHUJEN eikä mallina käytetä sananmuotoa ARVELUIDEN. Peruste valinnalle ei välttämättä ole tiedostettu, vaikka osaamme valinnan tehdä.

Esimerkkinä suomen kielestä otamme kaksitavuiset A-loppuiset subsstantiivit. Osaamme taivuttaa niitä, esim.:

KOIRA — KOIRIA	KAIRA — KAIROJA
ROTTA — ROTTIA	KANA — KANOJA
USVA — USVIA	KERMA — KERMOJA
KORVA — KORVIA	KIRVA — KIRVOJA

Näissä sanoissa havaitaan monikossa vartalonloppuisen A-äänteen joko katoavan tai muuttuvan O-äänteeksi. Useimmat meistä eivät kuitenkaan tunne sitä säännönmukaisuutta, jonka perusteella voidaan päätellä kummalla tavalla sanoja pitäisi taivuttaa.

Jos keksimme sellaisia kaksitavuisia A-loppuisia sanoja, joita emme entuudestaan tunne, kuten SEERA tai RUULA osaamme kuitenkin taivuttaa tällaisia. Luultavasti lukijakin muodostaisi monikkomuodot SEEROJA ja RUULIA (eikä SEERIÄ tai RUUULOJA).

Selitys löytyisi, jos suorittaisimme näitä taivutuksen kokeiluja vierustoverin nähden ja kuullen tai yksin ollessamme peilin edessä. Kun tarkkailisimme huulten asentoa niissä sanoissa, joista A häviää, huomaisimme hyvinkin sen, että aivan ensimmäisen vokaalin kohdalla suu olisi supussa. Sanan ääntämystä kuvaavilla termeillä sanoisimme, että A häviää, jos vartalon ensivokaali on ns. pyöreä vokaali (joita ovat suomessa O, Ö, U ja Y).

1.3 Kieli on moniselitteistä ja epätäsmällistä

Kieli ei ole aina ollut samanlaista, eikä kieli nytkään ole kaikille aivan samaa. Kielessä on myös moniselitteisyyt-

tä eri tasoilla. Yksittäisiä sananmuotoja tai virkkeitä voidaan tulkita useammalla tavalla. Kielen koodausjärjestelmät ovat löyhiä eivätkä luonnolliset kielet näiltä osin ole lainkaan ohjelmointikielten kaltaisia.

1.3.1 Kieli vaihtelee ja muuttuu

Kieli muuttuu. Sen havaitsee kyllä, jos lukee useamman kymmenen vuoden takaista tekstiä. Monet sanat, kuten DIREKTIIVI ja KÄNNYKKÄ eivät olleet tuolloin tunnettuja, ja silloinen teksti vaikuttaa kenties huolitellummalta kuin nykyinen. Jos tekstissä lukee HARAKKATA eikä HARAKKAA, voi arvata, ettei teksti ehkä ole aivan tuoretta (vaan esim. Juhani Ahoa). Jos taas tulee vastaan muoto SUTEA pro SUTTA, voi epäillä, että se on kirjoitettu lähiaikoina jossakin epämuodollisessa yhteydessä kuten keskustelupalstalla (”En oo koskaa nähny sutea luonnossa mut haluisin.”).

Olemassa olevat sanat voivat saada *uusia merkityksiä* olosuhteiden muuttuessa. Esimerkiksi ESTEETTÖMYYS kuvasi muutama vuosikymmen sitten sitä, että esimerkiksi asevelvolliselle voitiin myöntää passi, mutta tämän käytön jäätyä pois sana on vaivihkaa otettu uusiokäyttöön merkitsemään laitteen, ohjelman tai tilojen soveltuvuutta vammaisille.

Kielissä on myös alueellisia *murteita* ja toisaalta korkeasti koulutettu väki voi puhua ja kirjoittaa aika lailla eri tyylillä kuin vähemmän kouluja käynyt. Chatissä ja tekstiviesteissä käytetään ja kirjoitetaan kieltä vapaammin kuin pysyvämpään käyttöön tarkoitetussa tekstissä. Kaikki tämä vaihtelu ja muutos asettaa omia vaatimuksiaan

kieliteknologialle.

Puhuttu kieli on vaihtelevampaa kuin kirjoitettu. Kirjoitetulla kielellä on yleensä varsin *tarkat normit*, jotka usein nimenomaisesti määräävät kullekin sanalle yhden ainoan kirjoitusasun oikeaksi (esim. PAHOITTAA eikä PAHOTTAA) tai tietyille asioille yhden nimenomaisen suositellun ilmauksen. Puhekielille on aika lailla *erilaiset kielioppisäännöt* kuin kirjoitetuille kielille ja niillä on hieman eri lailla painotetut sanastonsa. Puhekielet ovat lauseja muoto-opillisesti selvästi erilaista kuin kirjoitetut. Vain harvojen kielten puhutun muodon kielioppeja on kuitenkaan vielä laadittu, vaikka hyvin monien jopa pienten kieliyhteisöjen kirjoitettujen kielten kieliopit ovat saatavilla.

Puhutussa kielessä käytetään helpommin paikallisia murrepohjaisia muotoja tai sanojakin. Myös tyyliä ja *huolittelun astetta* saatetaan vaihdella mielialan ja tilanteen mukaan. *Nopeasti puhuttaessa* tietyt kohdat paitsi lyhenevät niin myös yksinkertaistuvat ja joitakin osia sanoista voi jäädä kokonaan ääntymättä.

Osa puhutun kielen murrepohjaisista yksittäisten puhujien puheen eroista on *tiedostamattomia*. Usein puhuja luulee tietoisesti välttävänsä kaikkia kotimurteestaan muistuttavia piirteitä, muttei siinä kuitenkaan täysin onnistu. Kuitenkin harjaantunut ja murre-eroista tietoinen kuulija pystyy tunnistamaan tällaisia eroja.

1.3.2 Sananmuotojen moniselitteisyys

Sananmuotojen moniselitteisyyttä syntyy usealla tavalla. Toisaalta jo perusmuoto voi olla moniselitteinen sanaluokkansa tai merkityksensä puolesta. Toisaalta taivutus,

yhdyssanan muodostaminen tai sanajohto voi sattumalta aiheuttaa moniselitteisyyttä.

Esimerkiksi englannin kielessä moniselitteisyys on runsasta, ehkä noin puolet juoksevan tekstin saneista voitaisiin tulkita useammalla kuin yhdellä tavalla. Sananmuodot kuten HAND voivat olla joko substantiiveja (MY HAND WAS HURT) tai verbejä (PLEASE, HAND ME THAT PAPER).

Sanojen taivuttaminen voi tuottaa monitulkintaisuuksia eri tavoin. Suomen kielessä esim. HAUISTA syntyy substantiivista HAUKI yksikön elatiivina, substantiivista HAKU monikon elatiivina ja sanasta HAUIS yksikön partitiivina. Sijapääte -TA sisältyy sijapäätteeseen -STA, joten loppukirjainten puolesta elatiivimuodot voisivat olla partitiiveja jostakin toisesta sanasta. Lisäksi astevaihtelu voi muuttaa vartaloa toisen sanan kaltaiseksi, tässä pudottamalla HAUKI-sanan K-kirjaimen pois saadaan osa toisen sanan vartaloa. Yhteisvaikutuksena näistä ilmiöistä kieleen syntyy moniselitteisiä sananmuotoja.

Suomen kielessä saneista kuitenkin vain pieni osa on moniselitteisiä noin 10 %, mikä johtuu kai osittain siitä, että sanojen vartalot ovat melko pitkiä englantiin tai ruotsiin verrattuna. Monet suomen kielen taivutuspäätteet paljastavat, minkä sanaluokan sanasta on kyse.

Yhdyssana voi olla toisen sanan kaltainen, esim. sananmuodosta KUUTAMOILTA on vaikea arvata, tarkoitetaanko yhdyssanaa KUUTAMO-ILTA vaiko KUUTAMO-sanan monikon ablatiivia. Ruotsin kielessä sanojen vartalot ovat usein yksitavuisia ja saattavat alkaa useamman konsonantin yhdistelmällä tai vastaavasti päättyä siten. Tästä syntyy aika lailla periaatteellisia monitulkintaisuuksia kuten esim. sananmuoto FRUKOSTEN, joka voi-

daan tulkita paitsi yksiosaisena FRUKOST-sanan määrättynä muotona, myös useana erilaisena yhdyssanana kuten FRU+KO+STEN tai FRU+KOST-sanan määrättynä muotona jne.

Sananmuotojen moniselitteisyys on aika petollista sikäli, että ihminen *ei kovinkaan hyvin havaitse* sitä. Ihminen arvaa useimmiten sananmuotojen oikean tulkinnan huomaamattaan asiayhteyden ja odotustensa perusteella. Esimerkiksi virkkeessä AHDIN LUO PÄÄSI KUUSI ALAMAISTA jokainen sane on moniselitteinen, mutta sen havaitakseen täytyy virkettä katsoa aivan toisella silmällä.

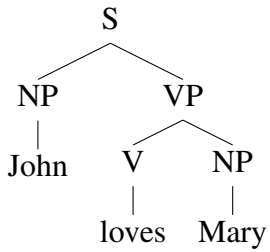
Koeta harjoituksen vuoksi itse päätellä seuraavien sananmuotojen kaikki tulkinnat (perusmuoto ja sanaluokka kustakin tulkinnasta): (a) ALUSTA, (b) KONEISTA.

1.3.3 Lauserakenteen moniselitteisyys

Saneista muodostuu lauseita ja lauseista virkkeitä, mutta mitä lauseiden rakenne oikein on? Yksittäisten saneiden rooleja lauseessa voidaan kuvata esimerkiksi seuraavasti: (jäljessä seuraavan substantiivin) genetiiviattribuutti tai adjektiiviattribuutti, (lauseita yhdistävä) alistuskonjunktio, (lauseen) adverbiaali, tai toisaalta esim. (lauseen) subjekti, objekti tai predikaatti. Ns. klassillinen lauseenjäsennys noudattaa tällaista kuvaustapaa, jossa kuhunkin yksittäiseen saneeseen liitetään sen *lauseopillinen funktio*.

Kielitieteellisissä teorioissa ja sittemmin myös tietokoneen ohjelmointikielissä on käytetty nk. *lausekerakennetta* (phrase structure), jota kuvataan usein puurakenteella.

Puurakenne on eräs tapa kuvata vierekkäisten saneiden tai niistä koostuvien laajempien kokonaisuuksien yhteen-



Kuva 1.2: Virkkeen ”John loves Mary” lausekerakenne puukuvaimena

kuuluvuutta. Puun osa, ns. alipuu edustaa lauseketta, jonka osat kuuluvat läheisemmin toisiinsa kuin alipuun ulkopuolella oleviin sanoihin. Kuvan 1.2 lausekerakenne ilmaisisi siten, että virkkeen JOHN LOVES MARY predikaatti LOVES ja sen objekti MARY kuuluvat tiukemmin yhteen kuin subjekti näihin kumpaankaan.

Eri kielissä noudatetaan erilaisia *sanajärjestyksiä* ja kielissä on myös erilaisia kieliopillisia keinoja lauseen rakenteen ilmaisemiseksi. Englannin kielessä saneen tai lausekkeen sijaintipaikka lauseessa on tärkeä keino, sillä irrallisista sananmuodoista ei juurikaan näkisi, ovatko ne verbejä, substantiiveja vai adjektiiveja, esim.:

THEY PAINT THE WALL

THIS PAINT IS WET

Paitsi yksittäisten saneiden sanaluokat, riippuvat englannissa saneiden roolit lauseenjäsennyksessä niiden keskinäisestä järjestyksestä, esim.:

THE ELEPHANT KILLED THE SNAKE

THE SNAKE KILLED THE ELEPHANT

Suomen kielessä puolestaan saneen sijainti lauseessa ei kerro kovinkaan paljon, mutta sanojen taivutuspäätteet sitäkin enemmän:

JÄNIS SÖI PORKKANOITA

PORKKANOITA SÖI JÄNIS

Sijamuodoilla, prepositioilla ja saneiden järjestyksellä tms. näkyviin merkitty lauserakenne ei kuitenkaan koodaa kaikkea, mikä tarvittaisiin merkityksen täsmällistä päättelemistä varten. Esimerkiksi ilmauksessa:

PUNAINEN TUPA JA PERUNAMAA

ei ole minkäänlaista näkyvää kieliopillista merkintää siitä, olisiko kyseinen perunamaa punainen vai ei. Ei perunamaa luultavasti punainen ole, mutta se tieto ei tule tuosta ilmauksesta vaan siitä, mitä maailmassa olevista asioista muuten tiedämme. Jossakin toisessa lauseessa kuten seuraavassa:

YLVÄS RYHTI JA KÄYTÖS

voimme hyvin arvata, että on kyse myös ylväästä käytöksestä, vaikka tästäkään ei ole näkyvää merkintää.

Aina ei oikea tulkinta selviä, kieliopillisten kriteerien, sanakirjan tai edes meitä ympäröivän fysikaalisen maailman tavanomaisilla ominaisuuksilla. Joskus tarvitaan lisänä yhteiseksi oletettua tietoa historiasta ja kulttuurista kuten seuraavassa:

SEN HÄN TEKI HYVÄLLÄ SYYLLÄ, SILLÄ CORTESIN NIMI OLI MEIDÄN PÄIVINÄMME YHTÄ KUULUISA KUIN CAESARIN ROOMALAISTEN KESKUUDESSA TAI HANNIBALIN KARTHAGOLAISTEN PARISSA.

Caesarin nimi oli kuuluisa roomalaisten keskuudessa, mutta Cortesin nimi ei voinut olla kuuluisa Caesarin roomalaisten keskuudessa, vaan kyseessä on vertailu … YHTÄ KUULUISA KUIN CAESARIN NIMI OLI ROOMALAISTEN KESKUUDESSA TAI …. Tässä NIMI OLI -sanojen pois jättäminen houkuttelee lukijaa väärälle polulle virkkeen tulkinnassa. Tarpeettomaksi katsottujen osien poisjättöä kutsutaan kielitieteessä *ellipsiksi* ja ilmiö vaikeuttaa erityisesti kielen rakenteen automaattista tunnistamista.

Seuraavan virkkeen tulkinta on yhtä lailla pulmallinen:

KUN NÄEMME SURKEIDEN ALKUASUKKAIDEN VEREN PUNAAMAN KÄDEN KOHOAVAN PYYTÄMÄÄN TAIVAAN SIUNAUSTA ASIALLE, TUNNEMME TÄMÄ TEON YHTEYDESSÄ JOTAKIN INHON TAPAISTA.

Irrallinen lause ei sisällä tarpeeksi tietoa yksiselitteistä tulkintaa varten. Ollaksemme varmoja, tarvitsisimme tietoa niiden tapahtumien kulusta, joihin virkkeen tekstissä viitataan. Tässä tapauksessa käsi oli konkistadorin käsi, jonka alkuasukasparkojen veri oli punannut, sulutuksen avulla ilmaistuna ((SURKEIDEN ALKUASUKKAIDEN)

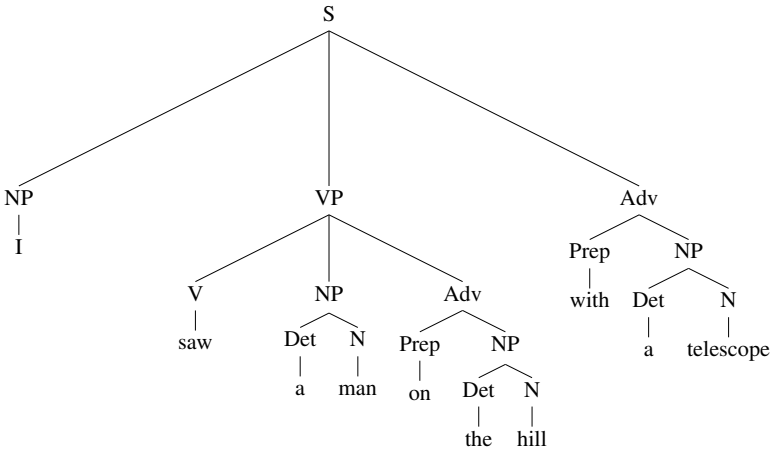
VEREN) PUNAAMAN KÄDEN. Kielioppi ei tässä auta sulkemaan pois toista tulkintaa, jossa kohoava käsi olisi ollut alkuasukkaiden (yhteinen) käsi, sulkujen avulla esitettynä (SURKEIDEN ALKUASUKKAIDEN) (VEREN PUNAAMAN KÄDEN). Rakenteellisesti kyse on siitä, että peräkkäiset genetiivimuotoiset substantiivit voidaan yhdistää erilaisissa järjestyksissä.

Kuvassa 1.3 on astetta aiempia mutkikkaampi englanninkielinen virke I SAW A MAN ON THE HILL WITH A TELESCOPE. Lauseen kääntäminen suomeksi riippuu olennaisesti siitä, millainen rakenne sille lähtökielessä hahmotetaan. Kääntäessä joutuu ottamaan kantaa siihen, oliko kaukoputki miehellä, minulla vai kukkulalla ja edelleen oliko mies, minä vai kaukoputki kukkulalla. Englannin kielen sanajärjestykseen perustuva syntaksi ei tee näiden eri tulkintojen välillä eroa, mutta merkityksen tulkinnan pohjana olevassa syntaktisessa rakenteessa ero pitää tehdä. Puhutussa kielessä oikean merkityksen voi usein tunnistaa äänenpainoista ja tauoista. Kuvassa 1.3 olevan rakenteen mukainen suomenkielinen vastine voisi olla: NÄIN KUKKULALLA OLEVAN MIEHEN KAUKOPUTKELLA.

1.3.4 Merkityksen moniselitteisyys

Edellä on puhuttu siitä, että tietty sananmuoto voi olla moniselitteinen, eli sananmuoto voi tyypillisesti olla kahden tai useamman eri lekseemin taivutusmuoto. Vaikka olisimme osanneet päätellä, mistä lekseemistä on kyse, ei moniselitteisyys lopu siihen.

Monet lekseemitkin ovat monimerkityksisiä eli hakusanalla on useampia alamerkityksiä. Otetaan esimerkik-



Kuva 1.3: Virkkeen ”I saw a man on the hill with a telescope” eräs lausekerakenne puukuvaimena

si suomen kielen LASKEA-verbi, jota voidaan käyttää erilaisissa merkityksissä: PÄÄSTÄÄ JOKU JONNEKIN, LIUKUA JOTAKIN ALAS, SUORITTAA ARITMETIIKKAA jne. Tämäkään moniselitteisyys, eli ns. polysemia ei ole kielen käyttäjälle kovin ilmeinen. Kussakin käyttöyhteydessä yleensä vain yksi näistä sanan alamerkityksistä tuntuu mahdolliselta, eivätkä muut tule edes mieleen, esim.:

KARJA LASKETTIIN LAITUMELLE.

HÄN LASKI JYRKÄN MÄEN ALAS YHDELLÄ SUKSELLA, MUTTA PYSYI PYSTYSSÄ.

LASKE NÄMÄ LUVUT YHTEEN.

1.3.5 Kieli on epätäsmällistä

Matemaattiset mallit voivat kuvata epävarmuutta ja epätäsmällisyyttä lukuarvoina, jotka kuvaavat tapahtuman *todennäköisyyttä* (eli lukuarvoa nollan ja ykkösen välillä) tai muuta suhdelukua. Todennäköisyyslaskenta ja tilastotiede antavat välineitä tällaisen epätäsmällisyyden hallitsemiseksi ja kuvailemiseksi. Todennäköisyyksien kanssa ei siis voida sanoa varmasti, kuinka tulee käymään, mutta voidaan laskea hyvinkin johdonmukaisesti ja tarkasti, kuinka keskimäärin tulee käymään, kun ilmiö toistuu monta kertaa.

Ihmisten kieli samoin kuin ihmisten ajattelu on aivan *eri tavalla epätäsmällistä*, kuin mihin mikään matemaattinen epätäsmällisyyden kuvaustapa on varautunut. Jos pohdimme vaikka yksinkertaisen substantiivin, TUOLI, merkitystä tai sitä, mitä voi kutsua tuoliksi, huomaamme, että tuoleja ovat kaikenlaiset istumiseen tarkoitetut huonekalut tai sellaisia muistuttavat. Käytännössä tuolit voivat olla monenlaisia, eikä niille voida antaa fysikaalista tai geometristä *määritelmää*. Muotoilijan luomus voi paljastua tuoliksi vasta, kun kuulemme selityksen. Totuttujen tuolien kaltaisuus on vielä subjektiivisempi ja sattumanvaraisempi käsite. Tällaisessa kielen epämääräisyydessä on kyse siitä, että käsitteet, kuten TUOLI ankkuroituvat toisiin käsitteisiin, kuten ISTUA ja HUONEKALU. Kielen käsitteet muodostavat tietynlaisen verkoston, jossa kukin käsite sijaitsee *suhteessa toisiin käsitteisiin*.

Myös sellaiset kielen käsitteet kuin HYVÄ tai NOPEA ovat *suhteellisia*, kelluvia ja vuosien kuluessa *muuttuvia*. Nopea tietokone ei esimerkiksi tarkoita samaa nyt kuin joitakin vuosia sitten, jos arvioimme nopeutta koneen yh-

den sekunnin aikana suorittamien käskyjen määrällä. Sen sijaan kunakin ajanhetkenä tavanomaista tehokkaampaa tietokonetta voidaan pitää nopeana.

Ehkä kielen epätäsmällisyyttä on toisinaan helpompi lähestyä vuorovaikutteisen keskustelun viitekehyksessä. Epämääräisyys ei useissa tilanteissa haittaa paljoa, koska melko viitteellinen vuorosana tai puutteellisesti ymmärretty ilmaus riittää silti viemään keskustelua hyvin eteenpäin. Keskustelun etenemistä mallinnetaan usein käyttämällä apuna puhujan ja kuulijan käsityksiä toisen osapuolen tiedoista ja uskomuksista sekä jotakin mallia keskustelun kohteena olevasta asiasta.

1.3.6 Ovatko moniselitteisyys ja epätäsmällisyys rakennevirheitä?

Olemme havainneet, että luonnollinen kieli on kovin moniselitteistä ja epämääräistä. Onko se virhe? Voitaisiinko ajatella, että kieli olisi näissä suhteissa ”korjattavissa”?

Jotkut asiat voisivat olla kai paremminkin, esimerkiksi keinotekoinen kieli esperanto on taivutukseltaan paljon yksinkertaisempi kuin pääosa luonnollisista kielistä (ellei peräti yksinkertaisempi kuin kaikki luonnolliset kielet). Kaikki esperanton substantiivit päättyvät O-kirjaimeen ja kaikkien sanojen taipuminen on säännöllistä, myös OLLA-verbin vastineen. Tällainen säännöllisyys helpottaisi joitakin kieliteknologisia tehtäviä ja vähentäisi toki myös moniselitteisyyttä. Sanojen taipumisen kuvaaminen täsmällisillä säännöillä ja jäsennysmenetelmillä onnistuu helpommin näin säännölliselle luonnolliselle kielelle. Kuitenkin juuri sanojen taipumisen kuvaaminen on se osa kielen

rakenteen automaattisesta tunnistamisesta, joka onnistuu luonnollisille kielille helpoiten.

Esperantossakin ja sen puoleen missä tahansa kuviteltavissa olevassa keinotekoisessakin kielessä on hakusanoilla alamerkityksiä. Esimerkiksi esperanton LUDI vastaa alamerkityksiltään jokseenkin englannin kielen PLAY-sanaa (soittaa, pelata, leikkiä, näytellä). Alamerkitysten ratkaiseminen lienee esperantossa jokseenkin yhtä vaikeata kuin muissakin kielissä. Syntaktiset moniselitteisyydet ovat varmaan läsnä ihmisen tietoisesti kehittämissä kielissä suunnilleen yhtä vakavina kuin alkuperäisemmissä luonnollisissa kielissä.

Moniselitteisyyksiä ei siten oikein pääse karkuun, vaikka vaihtaisi keinotekoiseen kieleen. Matemaattisten tai muiden ankaraa formalismia noudattavien kaavojen käyttö voisi olla yksiselitteisempää, mutta luultavasti niiden käyttö olisi ihmisille niin vaivalloista ja hidasta, että niistä tuskin on kilpailijoiksi luonnollisen kielen tekstille tiedon tallennusmuotona. Luonnollisen kielen moniselitteisyyttä ja epämääräisyyttä voidaan siten pitää enimmäkseen kielten hyödyllisenä ominaisuutena, jonka ansiosta kielet kehittyvät tarpeitten ja maailman mukana ja säilyvät käyttökelpoisina. Muuttuvassa maailmassa epätarkka ja moniselitteinen kieli toimii joustavasti ja hyvin, kun taas yksiselitteinen ja tarkka kieli helpommin lakkaa kokonaan toimimasta.

Luku 2

Kirjoittajan apuvälineet

Kirjoittajan apuvälineiden tehtävänä on auttaa kirjoittajaa eri tavoin. Vaikka nämä apuvälineet eivät kaiketi olekaan kieliteknologian vanhimpia sovelluksia, ne ovat varmaan kuitenkin niitä, joiden kanssa useimmat ensiksi joutuvat tekemisiin ja jotka eniten ovat helpottaneet päivittäistä työtä. Tässä luvussa käsitellään tekstin kirjoittajan apuvälineitä erityisesti oikeinkirjoituksen kannalta. Tuonnempana kerrotaan tiedonhaun, puheteknologian, kääntämisen ja kielenoppimisen välineiden yhteydessä niiden menetelmiin perustuvista työkaluista, jotka nekin voivat auttaa kirjoittajaa työssään.

2.1 Yleistä kirjoittajien apuvälineistä

Kun kirjoitamme, kirjoitamme tietenkin jotakin kieltä, eikä sitä voi kirjoittaa miten tahansa. Kirjoittamiseen liittyy joukko normeja esimerkiksi suomen kieltä kirjoitettaessa

mm. siitä,

- miten sanat tulisi kirjoittaa oikein (esim. KAAVOITTAA eikä *KAAVOTTAA), miten yhdyssanoja tulisi kirjoittaa yhteen (esim. SUOMENKIELINEN eikä *SUOMEN KIELINEN), erikseen tai väliviivan kanssa,

- miten välimerkkejä olisi käytettävä ja miten lyhenteitä käytetään (esim. SAK:N mutta NATON muttei NATO:N),

- miten isoja ja pieniä kirjaimia tulisi käyttää, millaisia ilmauksia pitäisi tyylisyistä karttaa sekä

- miten sanoja jaetaan rivin lopussa, jos kokonainen sana ei mahdu riville.

Ihmisten kirjoittama teksti ei aina noudata johdonmukaisesti kaikkia normeja. Poikkeamia pidetään yleensä virheinä ja syitä niihin on erilaisia. Virheitä syntyy tietämättömyydestä tai lipsahdusten kautta (ja joskus tietysti tietoisesti ja tahallaan). Kirjoittaja voi osata hyvinkin puhua jokseenkin moitteettomasti kieltään, mutta kirjoitetun kielen normit voivat silti olla hämäriä. Monet taas kirjoittavat vieraalla kielellä, jolloin kirjoittaja ei ole varma siitä, miten asia ilmaistaisiin luontevasti edes puhekielessä saati sitten kirjoitettuna.

Virheitä syntyy siis toisaalta vahingossa, esim. kun:

- sormi lipsahtaa näppäimistöllä oikean näppäimen ohi toiselle näppäimelle (esim. VIRHW),

- näppäin ei painukaan tarpeeksi syvälle niin, ettei aiottua kirjainta tulekaan (esim. VRHE tai sormi hipaisee tahattomasti jotakin muuta näppäintä niin, että tulee ylimääräinen kirjain VIRHGE),

- sormet toimivatkin eri järjestyksessä eli kirjain kiilaa toisen edelle (esim. VIHRE),

- ajatus keskeytyy ja virkettä jatkettaessa loppuosa ei olekaan johdonmukaista jatkoa alkuosalle tai kun

- leikataan ja liimataan virkkeen osia ilman, että muistetaan korjata päätteitä uuden järjestyksen mukaisiksi tai teksti jää muulla tavoin rikkinäiseksi.

Yleensä ollaan sitä mieltä, että suomen kieltä kirjoitetaan jokseenkin niin kuin äännetään ja sen vuoksi suomalaiset yleensä tietäisivät miten pitää kirjoittaa. Virheet tulisivat siten enimmäkseen näistä jälkimmäisistä eli tahattomista lähteistä. Aivan näin asia ei toki ole, vaan suomen kielen kirjoittamisessa on paljon sopimuksenvaraisia ja nimenomaisesti opittavia asioita, jollaisista yllä oli mainintoja.

Monissa muissa kielissä ääntämys ja kirjoitus ovat kauempana toisistaan kuin suomessa. Sellaisten kielten puhujien luku- ja kirjoitustaitokin on ehkä usein hatarampaa. Sellaisia kieliä kirjoitettaessa on luontevaa, että tullaan kirjoittaneeksi väärin myös siten, että kirjoitettu muoto kyllä kuulostaisi luettuna oikealta. Tekstiä näppäilleet sormet ovat silloin tuottaneet sen, mitä kirjoittaja aikoikin, mutta aikomus ei ole oikeinkirjoituksen ja kieliopin normien mukainen.

2.2 Oikeinkirjoituksen tarkistus ja korjaaminen

Oikeinkirjoituksen tarkistus auttaa löytämään vahingossa väärin kirjoitettuja saneita, esim. sellaisia, joissa sormi on erehtynyt ja yksittäinen kirjain on pudonnut pois, vääristynyt tai ilmaantunut liikaa. Yksinkertaisimmat korjausmenetelmät perustuvat olettamukseen, että useimmat tällä tavoin virheellisesti kirjoitetuista saneista eivät ole käytetyn kielen sananmuotoja lainkaan, ja ovat tunnistettavissa juuri tämän ominaisuuden perusteella. Tehtävä on vaikeampi, jos virheen tuloksena on syntynyt toinen sinänsä mahdollinen sananmuoto. Sellaisia voidaan tunnistaa tarkastelemalla niiden sijaintia kokonaisessa virkkeessä tai viereisten saneiden perusteella, mutta näitä menetelmiä käsitellään tuonnempana.

Oikeinkirjoituksen tarkistuksen tehtävän hahmottamiseksi käytämme johdannossa määriteltyjä kolmea erillistä termiä sanalle: lekseemi (eli hakusana), sananmuoto ja sane. Harjoittelemme hieman näiden käsitteiden käyttöä:

Tietyn tekstin sanaston laajuutta voidaan kuvailla laskemalla siinä käytettyjen lekseemien (eli hakusanojen) määrää. Vivahteikkaassa tekstissä on oletettavasti paljon eri lekseemejä. Olemme todenneet aiemmin, että suomenkielisestä substantiivilekseemistä KELLO saadaan taivuttamalla noin 2000 sananmuotoa, joiden joukossa ovat mm. KELLO, KELLON, KELLOSSA, KELLOSSANIKO, ... Jokin teksti kokonaisuudessaan voisi olla noin 56 000 saneen mittainen. Siinä tekstissä voisi olla 56 KUIN-sanetta (eli KUIN-sananmuodon esiintymää).

Näillä käsitteillä ilmaistuna siis tekstistä irralleen otet-

tua sanetta voidaan epäillä väärin kirjoitetuksi, jos se ei ole kielen minkään lekseemin kieliopinmukainen sananmuoto.

2.2.1 Yksinkertainen oikeinkirjoituksen tarkistus

Keräämällä suuresta määrästä tekstiä siinä esiintyvät sananmuodot, saa likimääräisen oikeinkirjoituksen tarkistimen. Tällaisella tarkistimella on eräitä etuja ja eräitä haittoja. Sananmuotojen luettelo on jokseenkin suoraviivainen tehdä ja seuraavassa esitettävä keino on näytettä sitä, mitä kieliteknologian eräillä muilla kursseilla opetetaan ja harjoitellaan. Esimerkiksi Unix- tai Linux-järjestelmässä muutaman rivin komennolla saa raa'asta tekstimateriaalista esille siinä olevat erilaiset saneet esiintymiskertoineen:

```
cat  kirja.txt |
  tr -d '0-9.,:;()=/"!+?<>'\' |
  tr 'A-ZÅÄÖ' 'a-zääö' |
  tr -s ' \t' '\012' |
  sort |
  uniq -c | less
```

Tämä muutaman valmiin ohjelman yhdistelmä käsittelisi `kirja.txt` -nimisessä tiedostossa olevan tekstin ensin siten, että (1) aluksi poistetaan numerot ja välimerkit, jonka jälkeen (2) isot ja pienet kirjaimet normalisoidaan pieniksi, (3) sananvälit muutetaan rivinvaihdoiksi, jolloin kukin sane on omalla rivillään, jonka jälkeen (4) rivit voidaan järjestää lajittelemalla nousevaan aakkosjärjestykseen ja lopuksi (5) yhdistää keskenään identtiset

rivit yhdeksi ja varustaa yhdistelyt samanlaisten rivien lukumäärällä. Tuloksena on pitkähkö lista, jonka osana voisi olla seuraavanlaista:

```
    ...
420 alkaa
  1 alkaahan
  1 alkaakaan
  7 alkaakin
  8 alkaako
    ...
```

Tällä menettelyllä saadaan luettelo, josta voidaan edelleen valita mekaanisesti esim. vähintään tietyn määrän kertoja esiintyneet sananmuodot tai sitten käydä lista lävitse tekstieditorin (eli tekstinmuokkaimen) kanssa, tarkastaa listassa olevat sananmuodot ja poistaa niistä virheellisiksi katsotut tai liian epätavallisina pidettävät sananmuodot. Loppu onkin tietotekniikkaa: kun tällainen luettelo koottu, se voidaan järjestää pienikokoiseksi ja tehokkaasti haettavaksi tietorakenteeksi, jota oikeinkirjoituksen tarkistusohjelma voi käyttää.

Sananmuotojen luetteloon perustuva menetelmä toimii kohtalaisen hyvin englannin kielelle, jossa lekseemeillä on vain vähän taivutusmuotoja, sananjohto on melko vähäistä ja yhdyssanatkin kirjoitetaan enimmäkseen erillisiksi saneiksi. Riittävän kokoisessa aineistossa on ainakin suuresta määrästä yleisimpiä lekseemejä kaikki tarvittavat sananmuodot mukana. Jos tarkistukseen liittyy virheiden korjausehdotusten tarjoaminen, on todellisesta aineistosta kerättyjen korjausehdotusten tarjoaminen turvallista,

koska ne ovat varmemmin kirjoittajalle tuttuja ja hyväksyttäviä sananmuotoja (kunhan aineisto oli tarkistettua ja virheetöntä).

Menetelmä ei toimi kovinkaan kelvollisesti sellaiselle kielelle, jossa on paljon taivutusmuotoja, produktiivinen yhdyssanan muodostus ja runsas sanojen johtamisen mahdollisuus. Tällöin nimittäin kyseisellä menetelmällä onnistutaan keräämään liian pieni osa kaikista mahdollisista ja teksteissä itse asiassa esiintyvistä sananmuodoista. Kun menetelmää sovelletaan uusiin teksteihin, jotka eivät olleet sananmuotojen keräilyn perustana, tulee vastaan uusia ja uusia aivan hyviä ja mahdollisia sanamuotoja. Oikeinkirjoituksen tarkistus lakkaa olemasta hyödyllistä, jos väärin epäiltyjä saneita on liiaksi. Muistamme edellisestä luvusta, että suomen kielen sananmuotoja on periaatteessa olemassa tähtitieteellinen määrä (siellä laskettiin 10^{24}).

2.2.2 Morfologiseen jäsentimeen perustuva oikeinkirjoituksen tarkistus

Morfologinen jäsennin (eli morfologinen analysaattori) on tietokoneohjelma, jolla on ns. leksikko, jossa sillä on tiedot kielen hakusanoista. Jäsentimen tehtävänä on etsiä annetulle sananmuodolle se lekseemi tai ne lekseemit, jonka tai joiden kieliopin mukainen taivutusmuoto kyseinen sanamuoto voisi olla. Esimerkkejä olemassa olevan suomen kielen morfologisen jäsentimen toiminnasta:

```
kellossa
   "kello"  N INE SG
kellossaniko
   "kello"  N INE SG 1SG kO
```

```
digitaalirannekellossa
   "digitaali_ranne_kello" N INE SG
myslikokojyvävälipalapatukkatehdas
   "mysli_koko_jyvä_väli_pala_patukka_tehdas"
    N NOM SG
katosta
   "katto"  N ELA SG
   "katos"  N PTV SG
katsta
```

Jäsentimen tehtävänä on siis löytää sananmuodolle, esim. KELLOSSA sellainen lekseemi, tässä KELLO, jonka kieliopillisesti hyväksyttävä muoto kyseinen sananmuoto on, ja samalla tulostaa muotoa kuvaavat morfosyntaktiset koodit N INE SG, jotka kertovat löydetyn lekseemin sanaluokan ja muodon, johon taivutettuna siitä syntyy etsitty sananmuoto. Moniselitteiselle sananmuodolle, kuten KATOSTA, löytyy useampi kuin yksi tulkinta. Väärin kirjoitetulle kuten KATSTA tai morfologisen jäsentimen sanaston ulkopuolelle jääneelle sananmuodolle ei löydy yhtään analyysiä.[1]

Morfologinen jäsennin tarjoaa siten mahdollisuuden suorittaa oikeinkirjoituksen tarkistusta. Enintä osaa jäsentimen analyysin tuloksesta ei kuitenkaan tarvita: hakusanalla ja taivutusmuotoa kuvaavilla koodeilla ei juurikaan ole käyttöä. Ainoastaan tieto siitä, että ainakin yksi tulos löytyi, on tässä tarpeellinen. Jos morfologista jäsennintä

[1] Useita FIN-CLARIN-hankkeeseen kuuluvassa HFST-hankkeessa laadittuja vapaasti käytettäviä morfologisia jäsentimiä on saatavissa verkosta ja kokeiltavissa osoitteessa: http://www.ling.helsinki.fi/kieliteknologia/tutkimus/finnwordnet/demot.shtml

käytetään oikeinkirjoituksen tarkistamiseen, on tavallista, että siitä tehdään rinnakkainen, riisuttu versio, jossa näitä perusmuotoa ja taivutusmuotoa koskevia osia ei ole. Riisuttu versio voi mahtua pienempään tilaan ja olla kenties nopeampikin kuin täysimittainen jäsennin.

2.2.3 Oikeinkirjoituksen tarkistuksen arviointia

Oikeinkirjoituksen tarkistuksen onnistumista voidaan arvioida seuraavilla kahdella suureella, joita käytetään erityisesti myös tiedonhaun puolella:

saanti, joka kuvaa sitä, kuinka suuri osuus etsityistä (eli tässä väärin kirjoitetuista saneista) löydetään (eli on algoritmin epäilyttäväksi merkitsemien saneiden joukossa) ja

tarkkuus, joka kuvaa sitä, kuinka suuri osuus löydetyistä (eli tässä väärin kirjoitetuiksi epäillyistä) on todella väärin kirjoitettuja saneita ja jotka algoritmin toivotaan löytävän..

Voidaan perustellusti väittää, että sananmuodostukseltaan rikkaissa kielissä (kuten suomessa) morfologiseen jäsentimeen perustuva oikeinkirjoituksen tarkistus tunnistaa suuremman osuuden oikein kirjoitetuista saneista kielen mahdollisiksi sananmuodoiksi kuin sananmuotojen listaan perustuva, koska morfologisen analyysin avulla voidaan tunnistaa ohjelman leksikkoon sisältyvistä lekseemeistä niiden kaikki taivutusmuodot sekä mahdollisesti myös yksinkertaisista sanoista muodostetut yhdyssanat ja johdetut

sanat ja niiden eri taivutusmuodot. Morfologisen analyysin tarkkuus kirjoitusvirheiden löytämisessä olisi luultavasti parempaa kuin pelkkään aineistoista kerättyyn sananmuotojen listaan perustuva saanti, koska sananmuotolistaan perustuva menetelmä merkitsee huomattavasti enemmän saneita epäilyttäväksi.

Ei ole aivan ilmeistä, kummalla menetelmällä saavutettaisiin parempi saanti kirjoitusvirheiden etsimisessä eli kumpi löytäisi suuremman osuuden todellisista väärinkirjoituksista. Jos sananmuotojen lista on virheetön, se hyväksyy vain todellisia eli oikeita sananmuotoja, mutta niin tekee morfologiseen analyysiinkin perustuva ohjelma.

Kumpaakin menetelmää vaivaa se, että melko usein todellisuudessa väärin kirjoitettu sane on kuitenkin kelvollinen kielen sananmuoto. Morfologiselle jäsentimelle tuskin on tässä suhteessa etua siitä, että voi kattaa huomattavasti suuremman määrän erilaisia sananmuotoja kuin sanelistaan perustuva. Tällaisten virheiden tunnistamiseksi tarvitaan kehittyneempää kalustoa, joka käyttää hyväkseen saneiden esiintymien laajempaa ympäristöä, joista tuonnempana.

Huono kattavuus merkitsisi sitä, että isompi osa todellisista virheistä jää löytymättä. Käyttäjä ei siitä työskennellessään juuri kärsi, mutta tuloksena syntyvän tekstin laatu tietysti kyllä. Huono tarkkuus sen sijaan on niin häiritsevää, että käyttäjä luultavasti luopuu moisesta apuvälineestä, jos tarkistusohjelma pysähtyy kovin usein ja enimmäkseen turhaan.

Hyvin tehtynä sanetasolla toimivat oikeinkirjoituksen tarkistusohjelmat ovat kuitenkin varsin käyttökelpoisia. Tällaisia tarkistusohjelmia on laajalti käytössä. Tunnis-

tamatta jäävät sananmuodot eivät välttämättä ole virheitä. Jotkut niistä voivat toistua useita kertoja. Siksi niitä voidaan käyttäjän valinnan mukaan lisätä järjestelmään, jolloin esim. oman organisaation ja asiakkaiden nimet ja lyhenteet eivät jatkuvasti vaadi huomiota. Käyttäjän ei oleteta lisäävän kaikkia tunnistamatta jääviä sananmuotoja sanakirjaan, sillä selvästi kertakäyttöisten ilmausten tai nimien lisäämisestä ei ole hyötyä. Päinvastoin, kovin laajaksi paisuvalla sanastolla on taipumusta hyväksyä runsaammin aitoja virheitä oikeina.

2.2.4 Väärin kirjoitettujen saneiden korjausehdotukset

Paitsi sitä, että kirjoittajan apuvälineiden tulisi tunnistaa kirjoittajan väärin kirjoittamat sanat, ohjelmalta voidaan myös toivoa korjausehdotuksia eli valistuneita arvauksia siitä, mitä kirjoittaja itse asiassa aikoi kirjoittaa tai mitä hänen pitäisi kirjoittaa.

Korjaustehtävä on eräissä suhteissa *vaativampi* kuin virheen paikallistaminen. Periaatteessa korjauksia ehdottavan ohjelman voi odottaa tarjoavan sellaisen tai sellaisia vaihtoehtoja, jotka tarkistusohjelma kelpuuttaisi ja jotka ovat ”mahdollisimman paljon” väärin kirjoitetun saneen kaltaisia. Käyttäjä tietysti odottaa näkevänsä ainoana (tai ensimmäisenä) ehdotuksena juuri sen sananmuodon, joka hänen piti kirjoittaa.

Itse asiassa ihmisen silmä on usein aika huono huomaamaan pitkissä saneissa olevia kirjoitusvirheitä, varsinkaan niiden keskellä olevia. Käyttäjä saattaisi epäillä, että ohjelma vain ei tunnista sanetta sen harvinaisuuden takia

ja lisää sen poikkeussanojen luetteloon. Mutta nähdessään väärin kirjoitetun saneen rinnalla sen korjatun muodon, käyttäjä kyllä oitis huomaa ja tunnustaa virheensä.

Oikeinkirjoitusta korjaavat ohjelmat etsivät tyypillisesti yhden tai useamman oikein kirjoitetun vaihtoehdon ja pyrkivät asettamaan todennäköisimmän korjausehdotuksen ensimmäiseksi. Kriteereinä ohjelmat voivat käyttää mm.:

- sitä, kuinka isoja muutoksia tarvittaisiin, jotta ehdotetusta sananmuodosta tulisi tekstistä löytynyt, oletettavasti väärin kirjoitettu sane, (usein oletetaan että yhteen saneeseen ei yleensä tule monia virhelyöntejä),

- eri vaihtoehtojen keskimääräisiä yleisyyksiä teksteissä, sillä eri sananmuotojen esiintymistodennäköisyyksissä on suuriakin eroja, sekä

- saneen lauseopillista ympäristöä, jonka mukaan jotkut vaihtoehdot ovat toisia todennäköisempiä.

Käytössä olevat oikeinkirjoituksen tarkistusohjelmat useimmiten tarjoavat varsin hyviä korjausvaihtoehtoja, mutta vaihtoehtojen paremmuusjärjestyksen päättelyä niissä tuskin on vielä kehitetty loppuun saakka.

Morfologiseen jäsentimeen perustuviin korjausehdotuksiin liittyy eräs ilmiö, joka erottaa ihmisen ja koneen kykyä tunnistaa luonnollista kieltä. Ihminen käyttää mieluusti hyväkseen saneen asiayhteyttä ja on jonkin verran huono tunnistamaan monimutkaisempia taivutusmuotoja irrallisina. Morfologiselle jäsentimelle taas asiayhteydestä

ei ole apua, eivätkä sille instruktiivit ole sen kummempia kuin genetiivitkään. Ihminen voi siten pitää eräitä ohjelman tarjoamia muotoja suorastaan väärin kirjoitettuina, kunnes joku kertoo, mistä sanasta ja muodosta todellisuudessa on kyse, esim. ihmiselle PAHAISTA voisi edustaa sananmuotoa PAHOISTA tai PARHAISTA vähän väärin kirjoitettuna pikemmin kuin PAHAINEN -sanan partitiivia.

2.3 Oikeakielisyyden ja kieliopillisuuden tarkistus

Kieliopintarkistusohjelmalla on tehtävänä löytää ennen kaikkea lauseyhteyteen liittyviä virheitä. Tällaisia epäjohdonmukaisuuksia voi syntyä esimerkiksi tekstinosia siirreltäessä tai vain lyöntivirheistä, joiden tuloksena on syntynyt toinen eli väärä, mutta sinänsä mahdollinen sananmuoto. Toinen käyttö kieliopintarkistusohjelmille on auttaa vajavaisesti vierasta kieltä osaavaa kirjoittajaa tuottamaan virheettömämpää tekstiä.

Syntaktisen jäsennysohjelman tehtävänä on tunnistaa virkkeistä niiden rakenne, johon katsotaan kuuluvaksi esim. saneiden keskinäiset määrityssuhteet (eli mikä on pääsana ja mikä määrite) ja roolit (kuten subjektina tai attribuuttina oleminen). Syntaktisen jäsennysohjelman soveltaminen oikeakielisyyden tai kieliopillisuuden tarkistamiseen on kuitenkin varsin ongelmallista, sillä käyttäjä ei oikeastaan hyödy tiedosta, että virkettä kokonaisuutena ei voida jäsentää, vaan hän tarvitsee tarkemmin kohdennettua palautetta.

Kieliopillisuuden ja oikeakielisyyden tarkistaminen onkin vaativa tehtävä kunnolla toteutettavaksi. Käyttökel-

poiset ohjelmat sisältänevät kahdenlaista materiaalia. Toisaalta ne pyrkivät *tunnistamaan tavanomaisiksi todettuja kielivirheitä*, mikä voi koostua yksittäisten kliseiden ja manerismien luetteloista, esim. kehotuksia välttää tietynlaisia fraaseja tai neuvoa korvaamaan huonoina pidettyjä termejä suositeltavilla. Tämä osa ei ole varsinaista jäsentämistä, vaan pikemminkin ei-toivottujen ilmausten tunnistamista ja niihin liitettyjen korjaus- tai tarkistuskehotusten näyttämistä.

Toinen puoli oikeakielisyyden tarkistamisessa voi koostua jäsentämisestä, mutta aivan *erityisestä tähän tarkoitukseen sovitetusta jäsennyksestä*, jonka tavoitteena on vain tunnistaa virkkeen sisällä olevia lausekkeita, ei niinkään laajempia yhteyksiä kuten rooleja lauseenjäseninä. Siinä missä normaali saksan kielen jäsennin edellyttäisi artikkelin ja pääsanan välistä suvun kongruenssia, kieliopillisuuden tarkistukseen sovitetun jäsentimen pitää olla sallivampi. Ensin se tunnistaisi substantiivilausekkeen välittämättä kongruenssista ja vasta sitten suorittaisi tarkistuksia. Tämä menettely antaa mahdollisuuden kohdentaa virhe mielekkäästi, jolloin käyttäjälle voidaan antaa helposti ymmärrettävää palautetta. Ohjelma voi esim. ilmoittaa että artikkeli tai adjektiivi ei ole siinä oikeassa suvussa, jota pääsana edellyttäisi.

2.4 Synonyymisanastot ja tesaurukset

Synonyymisanastolla voidaan joskus elävöittää tekstiä tai poistaa tautofonisia ilmauksia korvaamalla toisto sanan jollakin synonyymillä. Osa synonyymisanastoista osaa

löytää syötesanansa myös taivutusmuotoisena ja tarjoaa synonyymit vastaavassa muodossa, jolloin ne sopivat sellaisenaan tekstiin siirrettäväksi, esim. sananmuodolle KAUPASSA voisi ohjelma tarjota vaihtoehtoja PUODISSA, MYYMÄLÄSSÄ jne.

Kun jo aiemmin olemme maininneet morfologisen jäsentimen, voimme hahmotella sellaisen synonyymisanaston, joka soveltuisi suomen kaltaisille kielille, joissa sanat taipuvat.

1. Tunnistetaan tekstistä osoitetun saneen perusmuoto (tai perusmuodot) morfologisen jäsentimen avulla,

2. käytetään tätä perusmuotoa avaimena, kun etsitään tietokannasta synonyymejä ja

3. tuotetaan käänteisellä morfologisella jäsentimellä näin löydettyjen synonyymien alkuperäistä sanetta vastaavat taivutusmuodot.

Jotkut morfologiset jäsentimet ovat kaksisuuntaisia, siten että niillä voi myös tuottaa sananmuotoja, kun niille annetaan perusmuoto ynnä toivotun muodon taivutuskoodit. Yllä olevassa prosessissa nämä otetaan ykkösvaiheessa talteen ja liitetään kolmosvaiheessa synonyymien perusmuotojen perään ennen taivutusmuodon tuottamista.

2.5 Saneiden jakaminen rivin lopussa

Tekstin jakaminen riveille edellyttää tunnetusti joskus saneiden jakamista eli tavutusohjelmaa, jotta palstoista saadaan tasaisia ja tiiviitä. Epätasaisen oikean reunan ohella

kirjapainotaidon mukaan pidetään erityisen rumana sitä, että sanojen välit venyvät liikaa. Useimpien kielten kohdalla tällaiselle saneiden jakamiselle eli tavutukselle on olemassa selviä normeja, kuinka se pitää tehdä, vaikka normit ovatkin kielikohtaisia. Normien tarkoituksena on ohjata saneiden jakoa sellaiseksi, että lukija pystyy mahdollisimman hyvin hahmottamaan jaetun saneen oikein. Tämä saavutetaan eri kielissä eri tavoin.

Suomen kielessä tavujako myötäilee ääntämyksen mukaisia tavuja ja tavun rajalta jakaminen onkin useimmiten havaitsemisen kannalta hyvä. Esim. ruotsissa on kuitenkin kaksi erilaista tavujakojen periaatetta: toisaalta ääntämyksenmukainen eli fonologinen tavutus ja toisaalta morfologinen tavutus, jossa pyritään säilyttämään mahdollisimman ehjiä sanojen vartaloita edellisellä rivillä ja erottaa kokonaisia päätteitä, vaikka näin jaettaisiinkin muualta kuin äännettävien tavujen rajoilta.

Monissa kielissä, kuten englannissa on erityisiä tavutussanakirjoja, jotka määräävät sovinnaiset tai hyvinä pidetyt jakokohdat. Joissakin kielissä, kuten suomessa on pikemminkin joukko sääntöjä, joita tulee noudattaa.

2.5.1 Suomen kielen tavutussäännöt

Suomen kielen tavujakosäännöt voidaan luonnehtia seuraavasti:

1. Yhdyssanat jaetaan ensisijaisesti sananrajan kohdalta, esim. KANSAN-EDUSTAJA. Näin saatuja yhdyssanan osia jaetaan alempana olevien sääntöjen mukaan kuten itsenäsiä yksiosaisia sananmuotoja.

2. Yksiosaisissa saneissa (tai yhdyssanan osissa) noudatetaan ääntämyksenmukaisia tavurajoja, esimerkiksi: HUO-LES-TUT-TA-VIS-SA-KAAN. Suomenkielisissä sanoissa tavuraja sijaitsee yleensä CV-jakson edellä (C= konsonantti, V=vokaali). Tietyt vokaaliyhtymät muodostavat (pitkiä vokaaleja tai) diftongeja, joita ei jaeta. Muiden vokaalien välistä voidaan jakaa (vaikkakin vokaalilla alkavaa jälkiosaa pidetään rumana jakona).

3. Tavuun tarvitaan ainakin yksi vokaali, ei siis: ST-ROGANOFF eikä NAKKIST-ROGANOFF.

4. Sanan alusta tai lopusta ei saa erottaa yhden kirjaimen mittaista osaa, vaikka se olisi ääntämisessä tavu, ei siis: VALTI-O eikä VALTI-OMIES. Muutenkaan yhdestä vokaalista koostuvan tavun edeltä suoritettu jako ei ole hyvä, esim. ei: RADI-OTA.

5. Vierasperäisissä sanoissa ei saa rikkoa yhtä äännettä tarkoittavaa useamman kirjaimen yhdistelmää, ei siis esim.: RIC-HARD.

Näissä kriteereissä on kohtia, jotka eivät ole triviaaleja eivätkä ratkea pelkällä tavanomaisella ohjelmoinnilla. Yhdyssanojen jako ensin osiinsa on kieliteknologinen tehtävä ja se hoituu, jos meillä on käytettävissä morfologinen jäsennin. Jäsentäessään sananmuotoja, ohjelma tulee sivutuotteena tunnistaneeksi vartalot ja yhdyssanan osat. Näin sikäli, kun jäsenin tunnistaa sananmuodon. Väärin kirjoitetut jäävät yleensä tunnistamatta, mutta myös ennen näkemättömät erisnimet, ammattitermit tai muuten harvinaisten hakusanojen muodot.

Yksiosaisille sanoille edellytettiin yllä mainituissa säännöissä ääntämyksen mukaista tavujakoa. Tavutusohjelman pitäisi siis tietää, miten sanoja äännetään. Kotoperäisen sanaston osalta tehtävä näyttäisi olevan hallittavissa siltä osin, mikä vaikuttaa tavurajojen sijaintiin, kunhan pidetään lukua konsonanteista ja vokaaleista sekä erinäisistä kahden tai kolmen kirjaimen yhdistelmistä.

Vierasperäisten sanojen kohdalla tehtävä on kuitenkin vaikea, koska sanan alkukielen tai alkukielen mukaisen ääntämyksen päättely voi olla valistuneellekin ihmiselle joskus epävarmaa ja edellyttää joka tapauksessa monen kielen ääntämyksen sääntöjen ainakin summittaista tuntemista.

Onneksi kuitenkin kirjainten yhdistelmät kertovat usein siitä, mitkä sanat ovat esim. sivistyssanoja tai lainasanoja. Tätä tietoa voidaan hyödyntää suomen kielen säännöistä poikkeavien tavurajojen tunnistamiseen, erityisesti etuliitteiden ja yhdyssanojen erottamiseksi oikein, esim. sanan-muodossa YÖ-KLUBI kirjainyhtymä KL on suomen kielessä harvinainen, mutta lainasanoissa yleensä aloittaa sanan (ja tämä tieto auttaa välttämään jakamisen konsonanttiyhtymän välistä).

Silloin kun tehdään tavutusalgoritmia ilman morfologista jäsennintä (tai jaetaan saneita, joita jäsennin ei tunnista), on vaara tehdä virheitä sananrajan tienoolla: esim. KANSA-NE-LÄKE tai PARIO-VI. Näitä voidaan välttää varomalla jakoja tietyissä ympäristöissä, esim. tyyppiä VNVCV olevilla alueilla. Niistä huomattava osa on genetiivialkuisia yhdyssanoja, joissa N-kirjaimen edestä jakaminen olisi virhe, mutta muulloin taas juuri se olisi oikea kohta. Varovaisuus ja jakokohtien välttäminen voi olla

viisasta paikoissa, joissa virheen mahdollisuus on merkittävä.[2]

2.6 Luettavuuden arviointi

Erilaisiin tarkoituksiin olisi hyvä saada nopeasti ja helposti arvioita siitä, onko kirjoitettu teksti helppolukuista vai vaikeaa. Tekstin luettavuutta on yritetty arvioida eri tavoin. Varhaiset ohjelmat olivat hyvinkin yksinkertaisia ja arvioivat luettavuutta esimerkiksi *virkkeiden tai saneiden pituuksien* perusteella. Tällaiset kriteerit korreloivat jossain määrin tekstin vaikeuden kanssa, mutta tekstin ymmärtämisen helppous tai vaikeus ei kuitenkaan suoraan riipu tällaisista kriteereistä. Tekstissä voi olla keskimäärin pitkiä virkkeitä ilman, että se on vaikeaa. Pitkät saneet eivät sinänsä ole vaikeita (vaikka ehkä englannissa latinalaisperäiset sanat ovatkin pitempiä kuin kielen omat tavalliset sanat).[3]

Yhtenä kriteerinä voisi käyttää tekstin *sanaston laajuutta* tai käyttää apuna tietoa sanojen keskimääräisestä esiintymisestä kielessä. Teksti voisi olla vaikeaa, jos siinä on tavanomaista enemmän harvinaisia sanoja.

Lukija arvannee, että sanaston laajuuden arviointi onnistuu englannin kielen tapauksessa aika hyvin tarkkailemalla saneiden ja erilaisten sananmuotojen osuuksia. Suo-

[2]Tavutusohjelmia ja oikeinkirjoituksen tarkistusohjelmia on kokeiltavissa verkossa eri osoitteissa, mm. `http://www.lingsoft.fi/demot`.

[3]Avoimen lähdekoodin toteutus tällaisista työkaluista on saatavissa mm. GNU-hankkeen sivuilta: `https://www.gnu.org/software/diction/`

men kielelle tämä kriteeri olisi laskettavissa paremmin, jos käytämme morfologista jäsennintä. Johto-opin hallitseva jäsennin varmaan myös tarjoaisi mahdollisuuksia tarkemmin erottaa tyyliltään raskaita sananmuotoja, kuin pelkkä sananmuotojen pituuden laskeminen.

On houkuttelevaa ajatella, että samantapaisella pintapuolisella *lauseenjäsennyksellä*, joka on hyödyllistä oikeakielisyyttä tarkistavassa ohjelmassa, olisi käyttöä tekstin luettavuuden arvioinnissa. Tällaisen avulla saataisiin esille suureita, jotka toivon mukaan paremmin kuvaisivat sitä vaikeutta, joka lukijalle tekstistä aiheutuu. Spekulaatioiden sijaan pitäisi kuitenkin tehdä selvityksiä ja käytännön kokeita, koska sitä, millaiset rakenteet ihminen kokee vaikeiksi ymmärtää ja millaiset helpoiksi, ei tunneta vielä kovinkaan hyvin.

2.7 Kirjoittajan apuvälineiden toteutusten teknologioita

Morfologisen jäsentimen tai tavutusalgoritmin laatiminen on toki eräänlainen ohjelmointitehtävä. Sen tekee erityiseksi kuitenkin se, että kohteena on luonnollinen kieli. Muistamme, että kieli ei ollut pieni, mistä seuraa että tehtävät voivat olla tavallista vaativampia ohjelmoitavaksi. Toisaalta kielen säännönmukaisuudet operoivat hiukan toisenlaisilla käsitteillä kuin ohjelmointi yleensä, esim. vokaaleilla, ei-pyöreillä vokaaleilla, soinnillisilla konsonanteilla tai diftongeilla ja umpitavuilla. Tästä seuraa lisää vaatimuksia ohjelmoinnille. Numeerisilla arvoilla on paljon vähemmän tekemistä kieliteknologiassa kuin ohjelmissa yleensä.

2.7.1 Äärellistilaiset automaatit

Apua löytyy mm. *äärellisten automaattien* ja *transduktorien* teknologiasta, jonka avulla avotavuja tai diftongeja jne. voidaan helposti ilmaista säännöllisinä lausekkeina. Äärelliset automaatit ovat periaatteessa hyvin yksinkertaisia ja hyvin ymmärrettyjä, melkeinpä kaikkein yksinkertaisimpia abstrakteja koneita, jotka pystyvät tekemään edes jotakin. Kuitenkin niiden avulla on suoraviivaisesti ilmaistavissa esim. suomen kielen kaikkien mahdollisten sananmuotojen joukko, astevaihtelu tai vokaalisointu. (Näistä asioista enemmän kieliteknologian aineopintojen kursseilla).

Oikeinkirjoituksen tarkistin suorittaa pohjimmiltaan tunnistustehtävää, jossa vastataan kysymykseen: ”Kuuluuko annettu sane eli merkkijono ennalta tunnettuun kielen kaikkien sananmuotojen joukkoon?” Tällainen tunnistaminen voitaisiin tehdä laatimalla tavanomaisia tietokoneohjelmia, jotka aikansa laskettuaan antaa tulokseksi ”kyllä” tai ”ei” (eli 1 tai 0). Haittana suoraviivaisessa menettelyssä on se, että jokaiselle kielelle pitäisi jokseenkin erilainen ohjelma. Lisäksi kerran laaditun ohjelman päivittäminen ja edelleen kehittäminen saattaa olla hyvinkin vaikeata, jos esim. päättelyiden keskinäistä järjestystä haluttaisiin muuttaa.

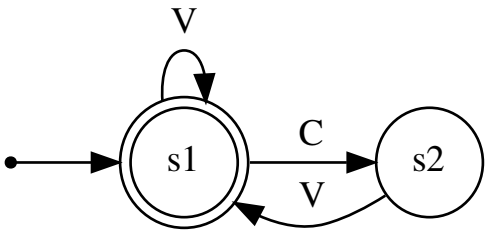
Paremmaksi lähestymistavaksi on osoittautunut tarkoitukseen sopivien formalismien käyttäminen. Morfologisen jäsentimen laatimiseksi ajatellaan yleensä tarvittavan tapa esittää sanavarasto eli leksikko sekä tapa esittää sääntöjä, joiden avulla sananmuotoja voidaan tuottaa. Kumpaakin varten tarvitaan jokin esitysmuoto, joka soveltuu kohteeseensa ja keinot näin esitettyjen sanojen tai sääntöjen to-

teuttamiseksi jollakin riittävän tehokkaalla tavalla.

Äärelliset automaatit perustuvat hyvin yksinkertaiseen tilasiirtymämekanismiin, joka on helppoa toteuttaa tietokoneella. Toisaalta äärellisiä automaatteja voidaan kuvailla ns. säännöllisillä lausekkeilla, joiden avulla monia kielen kannalta tärkeitä asioita voidaan ilmaista helposti. Äärellisten automaattien avulla voidaan helposti esittää mm. sananmuotojen luetteloita. Tällaiset luettelot voidaan pakata pieneen tilaan ja silti tällaisesta rakenteesta voidaan tehokkaasti ja nopeasti tarkistaa, onko annettu sananmuoto etukäteen annetussa joukossa. Näin voitaisiin esimerkiksi aiemmin mainittu yksinkertaistettu oikeinkirjoituksen tarkistusohjelma toteuttaa, ja monet niistä on myös käytännössä näin toteutettu.

Äärellinen automaatti koostuu rajallisesta määrästä *tiloja*, joita on tapana kuvata ympyröillä. Eräs niistä nimetään *alkutilaksi*, jossa automaatti on alussa. Automaatin tehtävänä on *tunnistaa tai hylätä merkkijonoja*. Kunkin merkin kohdalla pitäisi automaatista löytyä *tilasiirtymä*, joka on nimiöity kyseisellä merkillä, jolloin automaatti lukee ja kuluttaa kyseisen merkin ja siirtyy siirtymän osoittamaan uuteen tilaan. Jos automaatti viimeisen merkin jälkeen on jossakin ns. *lopputiloiksi* merkityistä tiloista, sen sanotaan hyväksyvän kyseisen merkkijonon. Hylätyksi merkkijono tulee, jos sitä tunnistettaessa jossakin kohdassa ei ole yhtään sopivaa siirtymää, tai jos se lopuksi on tilassa, joka ei ole lopputila.

Äärellistilaiset *transduktorit* poikkeavat äärellisistä automaateista siten, että ne paitsi hyväksyvät tai hylkäävät syötettyjä merkkijonoja, ne myös muuntavat hyväksytyt jonot toisiksi merkkijonoiksi. Morfologiset jäsentimet voi-



Kuva 2.1: Äärellistilainen automaatti, joka hyväksyy avotavuista koostuvat sanat ja hylkää ne, joissa on yksikin umpitavu (V=vokaali, C=konsonantti)

daan toteuttaa äärellisinä transduktoreina, jotka ovat äärellisiä automaatteja jonkin verran yleisempiä mekanismeja. Jäsennin muuntaa saneen sitä vastaavaksi perusmuotoiseksi hakusanaksi ynnä sanetta vastaavaksi taivutustiedoksi tilasiirtymien mekanismin avulla.

Synonyymisanasto voidaan hahmottaa myös siten, että siinä on kaksi morfologista jäsennintä yhdistettynä toisiinsa. Ensimmäinen jäsennin tunnistaa sitä sananmuotoa, jota käyttäjä osoittaa ja tuottaa siitä perusmuodon ja taivutuskoodit. Varsinainen synonyymitieto on synonyymien perusmuotojen muodostamissa ryhmissä, joiden avulla näin syntyneeseen perusmuotoon liitetään vuoron perään synonyymiryhmän muut perusmuodot, mutta kunkin perään pannaan samat taivutuskoodit, jotka alkuperäistä sanan-

muotoa analysoitaessa saatiin. Näin syntyneet perusmuoto ynnä taivutuskoodit -yhdistelmät syötetään takaperin käännettyyn morfologiseen jäsentimeen, joka siis tuottaa perusmuodon ja koodien perusteella sananmuodon. Äärellisistä automaateista ja transduktoreista sekä tämänlaisista menetelmistä puhutaan enemmän eräillä kieliteknologian aineopintojen kursseilla. Äärellistilaisten transduktorien soveltamisesta morfologiseen jäsentämiseen on kirjoitettu kattava ja hyödyllinen oppikirja (Beesley & Karttunen, 2003).

2.7.2 Toisinkirjoitusmekanismit

Tavujako-ongelmassa voidaan käyttää ensi vaiheena morfologista jäsennintä, jolloin tunnistetaan yhdyssanojen rajat. Sopivilla ohjelmilla voidaan soveltaa sanahahmoja koskevia sääntöjä loppujen jakokohtien tunnistamiseksi. Näitä sääntöjä voidaan toteuttaa joko äärellisillä transduktoreilla tai esim. Beta-nimisen ohjelman toisinkirjoitussäännöillä (Karlsson & Koskenniemi, 1990).

Luku 3

Tiedonhaku ja siihen liittyvät sovellukset

Tietokoneen muodossa saatavilla olevan varsinkin Internetissä sijaitsevan luonnollisen kielen tiedon määrä on kasvanut valtavaksi. Tietojen etsimiseksi on nykyään tehokkaita välineitä tarjolla ja ne ovat muodostuneet suurten joukkojen jokapäiväisiksi työkaluiksi. Paitsi kykyä tallettaa valtavia määriä dokumentteja tai niiden kopioita sekä saada yksittäisiä dokumentteja esille, tarvitaan monenlaisia muitakin menetelmiä ja työkaluja, jotta pärjättäisiin suunnattomien tietomäärien kanssa:

- **Tiedonhaku** eli dokumentteja (tai verkkosivuja) on haettava entistä valikoivammin ja löytyviä sivuja on voitava asettaa paremmuusjärjestykseen (eli relevanssijärjestykseen).
- **Tiedon tiivistäminen** eli pitkistä teksteistä pitäisi voida osoittaa sellaisia kohtia, joista käy ilmi teks-

tin pääasia.

- **Dokumenttien luokittelu** eli viesteistä tai teksteistä pitäisi päätellä minkä lajisia ne ovat, esim. kenen käsiteltäviksi ne kuuluvat vai ovatko ne kokonaan asiaankuulumatonta (siis roskapostia).

- **Indeksointi** eli teksteistä pitäisi voida tehdä asiahakemistoja joko koneen avustamana tai automaattisesti.

- **Semanttinen web** eli sisältöä pitäisi voida linkittää hienojakoisemmin eli kunkin alan määrämuotoisen käsitteistön avulla.

- **Automaattinen hypertekstin muodostaminen** eli dokumentteja pitäisi linkittää toisiinsa tunnistamalla kohtia, joissa sanotaan tärkeitä asioita, joihin muualta kannattaisi viitata, sekä toisaalta kohtia, joissa olisi tarvetta viitata sellaisiin paikkoihin, joissa siinä käsitellystä asiasta on lisää tietoa.

3.1 Tiedon haku

Tietoja on verkossa ja erilaisissa tietojärjestelmissä nykyään enemmän kuin tarpeeksi. Aikoihin ei yleisimpänä pullonkaulana ole ollut se, ettei tietoa ole olemassa, vaan vaikeutena on joko se, että haut tuottavat liikaa tuloksia tai että on vaikeaa löytää oikea dokumentti, koska siinä ehkä ei ole juuri niitä sananmuotoja, joilla etsittiin. Yksinkertaisesti sanottuna tiedonhaulla tarkoitetaan niiden dokumenttien löytämistä, jotka olisivat RELEVANTTEJA haun kannalta eli sisältäisivät sitä tietoa, jota etsitään.

Tiedonhaun onnistumisen arvioimiseen liittyy kaksi tärkeää suuretta, joista oli jo puhetta oikeinkirjoituksen tarkistamisen kohdalla:

- **saanti** (engl. recall), joka kuvaa sitä prosenttiosuutta relevanteista (eli halutuista) dokumenteista, joka kaikista relevanteista dokumenteista tuli esille haun avulla ja
- **tarkkuus** (engl. precision), joka kuvaa sitä kuinka suuri prosenttiosuus haun esille tuomista dokumenteista oli relevantteja.

Haun tuloksena *hakukone* etsii tietyn joukon dokumentteja. Saanti kuvaa siis sitä osuutta koko aineiston kaikista relevanteista dokumenteista, joka tuli tähän haun tulokseen mukaan. Yleensä osa jää syystä tai toisesta löytymättä. Toisaalta hakutulokseen eksyy myös muitakin kuin haluttuja dokumentteja eli hakutuloksemme ei koostu pelkästään relevanteista dokumenteista, vaan mukana on roskaakin. Käyttäjä toivoisi luonnollisesti, että haku onnistuisi 100 % saannilla ja 100 % tarkkuudella, mutta käytännössä tällaiseen ei yleensä voida päästä, vaan on tyydyttävä kompromisseihin. Realiteettien puitteissa ei ole edullista tavoitella vain toisen kriteerin maksimointia, sillä sisällyttämällä aineiston aivan kaikki dokumentit haun tuloksiin, saadaan 100 % saanti, mutta aivan mitätön tarkkuus. Samoin, tyytymällä vain yhteen todennäköisimpään dokumenttiin saavutettaisiin usein 100 % tarkkuus mutta mitätön saanti. Erikoistapauksissa voidaan toki haluta hakukoneelta vain yksi todennäköisin osuma.

Relevanssi voi olla eriasteista ja sen vuoksi tiedonhaun ohjelmien pitäisi mieluusti tarjota dokumentteja siten, että

relevantimmat tulevat ensiksi. Usein ajatellaan, että käyttäjä hakee oikeastaan kysymykseensä parhaiten vastaavaa yhtä parasta dokumenttia ja hän haluaa löytää sen hakutulosten kärjestä. Tällöin hakutulosten priorisointi sopivan *relevanssikriteerin* avulla on hyvin tärkeää.

Otetaan kuitenkin vielä esimerkki saannin ja tarkkuuden laskemisesta. Oletetaan, että tekstitietokannassa olisi yhteensä 10 000 dokumenttia (millä ei itse asiassa ole saannin ja tarkkuuden laskennassa väliä). Näiden joukossa olisi 30 kalastamista käsittelevää dokumenttia. Niiden löytämiseksi etsimme kriteerillä KAL- eli saneita, jotka alkavat noilla kolmella kirjaimella. Tällä haulla saamme yhteensä 36 dokumenttia. Nämä eivät kaikki kuitenkaan ole kalastamista koskevia, ainoastaan 18 niistä on kalastamista koskevia eli halutunlaisia (ja lopuissa on muita KAL -alkuisia saneita, esim. KALLOSSA tai KALKKIA). Nyt voimme laskea saannin, joka on 18/30 eli 60 % sekä tarkkuuden, joka on 18/36 eli 50 %. Huomaamme, että relevanssin eri asteilla ei saannin ja tarkkuuden laskemisessa ole sijaa.

Itse haku suoritetaan yleensä dokumentissa esiintyvien *sananmuotojen* perusteella, siten että kelpuutetaan lähinnä vain sellaisia dokumentteja, joissa kaikki vaaditut sananmuodot (tai mahdollisimman monet niistä) esiintyvät. *Relevanssin* arviointiin sen sijaan käytetään erilaisia tilastollisia menetelmiä, jotka arvioivat lisäksi eri sanojen kykyä erotella dokumentteja toisistaan. Sanojen katsotaan erottelevan dokumentteja hyvin, jos niiden esiintymät keskittyvät osaan dokumenteista ja vastaavasti huonosti erottelevat sellaiset sanat, jotka esiintyvät tasaisesti kaikissa dokumenteissa. Sanan esiintyminen useasti tietyssä doku-

mentissa parantaa vielä kyseisen dokumentin pisteytystä relevanssin suhteen. Sananmuotojen esiintymiskertojen lisäksi relevanssiin vaikuttavat muutkin asiat kuten se, miten monet ja millaiset muut sivut viittaavat kyseiseen sivuun ja niinkin arkiset asiat kuin, esiintyvätkö saneet dokumentin alussa vai lopussa ja edelleen, esiintyvätkö otsikoissa vai juoksevassa tekstissä.

Dokumentteja etsitään useimmiten niissä esiintyvien sananmuotojen perusteella, koska tällainen on teknisesti helpointa toteuttaa. Kieliteknologialla on useampia mahdollisia tehtäviä tiedonhaun yhteydessä luonnollisen kielen ominaisuuksien takia.

1. Monissa kielissä hakua vaikeuttaa se, että sanat voivat esiintyä *useissa eri taivutusmuodoissa*. Useat hakukoneet etsivät vain tarkkaan annetuilla merkkijonoilla, jolloin perusmuodolla haettaessa muut taivutusmuodot jäävät löytymättä. Käyttäjän pitäisi luetella hakukriteerissä lekseemin kaikki taivutusmuodot, mikä on vähintäänkin hankalaa, jos muotoja on paljon. *Hakuavaimen katkaiseminen* lopusta on joissakin hakukoneissa mahdollista, mutta katkaisemiskohdan merkitseminen vaatii käyttäjältä tarkkaavaisuutta ja taitoa. Katkaistulla perusmuodolla tehty haku voi löytää muitakin kuin toivottuja sananmuotoja.

2. Kun dokumentissa puhutaan tietystä asiasta, se voidaan ilmaista eri sanoin kuin, mitä kyselijä tulee ajatelleeksi laatiessaan hakulauseketta. Dokumenteissa esiintyvien *synonyymisten ilmausten* takia haun saanti yleensä heikkenee.

3. Hakuavaimen sanoilla voi olla erilaisia *alamerkityksiä*, joista vain jokin liittyy haettavaan asiaan ja muut eivät. Itse haku on mekaaninen ja voi tuottaa runsaasti turhia eli irrelevantteja dokumentteja, jossa puhutaankin muusta, jolloin haun tarkkuus kärsii.

Jotkut Internetin hakukoneista (itse asiassa harvat kuten AltaVista) sallivat haut paitsi kokonaisilla myös *katkaistuilla sananmuodoilla*. Tällaisilla tyvillä voi tavoittaa taipumisen kautta tai muuten yhteen kuuluvia sananmuotoja (ja siten parantaa saantia), kunhan niissä on tarpeeksi paljon yhteistä alkuosaa (jotta tarkkuus ei liiaksi kärsi). Esim. prefiksillä KARAMANLI- löydettäisiin hyvin dokumentit, jotka puhuvat Karamanlis-nimisestä henkilöstä. Sen sijaan prefiksi KÄ- kyllä löytäisi KÄSI-sanan sisältävät dokumentit, mutta paljon muitakin.

Haut voivat koostua sananmuotojen luetteloiden sijasta myös ns. *Boolen lausekkeista*, joilla yksittäisiä hakutermejä yhdistellään mutkikkaammin. Boolen lausekkeilla voidaan esittää vaihtoehtoja hakutermeille (esim. LAIVA OR ALUS) ja mahdollisia kieltoja (esim. NOT MAINE). Boolen lausekkeita pidetään usein vaikeina tavallisen käyttäjän kannalta. Boolen lausekkeiden käyttö ei myöskään ole kieliteknologiaa vaan tavanomaista tietotekniikkaa.

Eksplisiittisen hakulausekkeen muodostamisen sijasta tiedonhakua voidaan toteuttaa myös siten, että käyttäjä pyytää toisia, *dokumentin kaltaisia dokumentteja*. Haun ohjaaminen tällä tavoin on käyttäjälle yleensä helpompaa kuin monimutkaisen hakulausekkeen muuntelu. Hakuohjelman tehtävänä on silloin pikemminkin arvioida kahden dokumentin samanlaisuutta esim. niissä esiintyvien sananmuotojen avulla, mikä puolestaan on selkeästi määritelty

laskentatehtävä. Samanlaisten dokumenttien hakua käytetään yleensä toisena vaiheena, kun käyttäjän antamalla hakulausekkeella ensin on haettu keskenään mahdollisimman erilaisia dokumentteja, joista käyttäjän on ehkä helppo valita parhaiten sopiva.

Internetissä on vapaasti käytettävissä erilaisia *hakupalveluita*.[1] Jotkut tarjoilevat etukäteen laaditun luokituksen mukaan viitteitä, jotkut etsivät viitteitä eri puolilta verkkoa kerättyjen sivujen sisältämien sananmuotojen perusteella.

Dokumentin sisältämien sananmuotojen perusteella hakeminen perustuu *hakurobottiin* eli ohjelmaan, joka vaeltaa ympäri koko Internetiä ja kerää sivuja. Löydettyään sivun se tallettaa sen ja jatkaa etsimistä kyseisen sivujen sisältämien uusien linkkien perusteella. Kaikista tällä tavalla löydetyistä verkkosivuista eristetään haun pohjaksi niissä esiintyvät sananmuodot. Läpikäynnissä löydettyjen sivujen sananmuodoista kerätään suuri *hakemisto*, jonka perusteella voidaan päätellä, millä verkkosivuilla esiintyy mitäkin sananmuotoja ja erityisesti toisinpäin, mitkä ovat ne sivut, joilla tietyt sananmuodot ja sananmuotojen yhdistelmät esiintyvät.

Tavallinen haku koostuu joukosta sananmuotoja, joiden kaikkien edellytetään esiintyvän haettavissa dokumenteissa ja vaativammissa hauissa Boolen lausekkeella määritellään dokumentissa vaadittavat sananmuotojen yhdistelmät. Hakukone löytää pyynnön perustella (toivottavasti kohtuullisen kokoisen) joukon dokumentteja, joiden seassa (toivon mukaan) pääosa parhaiten relevanteista verkkosivuista on.

[1] www.google.com, www.bing.com/, www.altavista.com, www.yahoo.com, jne.

3.2 Monikielinen tai kielten välinen tiedonhaku

Maailma ja Eurooppa ovat monikielisiä. Kouluja käyneet ihmiset osaavat usein muutakin kieltä kuin äidinkieltään. Toisen kielen taito ei kuitenkaan yleensä ole äidinkielen veroinen. Osataan kyllä lukea, eli jos dokumentti löytyy, siitä on hyötyä vieraskielisenäkin. Ongelma voi olla siinä, että hakija ei osaisi *ilmaista hakukriteeriä* vieraalla kielellä läheskään yhtä onnistuneesti kuin äidinkielellään.

Toinen monikielisen tiedonhaun (cross-language information retrieval, CLIR) tehtävä perustuu yksinkertaisesti siihen, että olisi vaivalloista laatia erikseen hakulauseke *kaikilla kyseeseen tulevilla kielillä*. Kone voisi tässä auttaa ratkaisevasti muuntamalla haun annetusta kielestä toisillekin kyseeseen tuleville kielille sopivaksi käyttäen apunaan esimerkiksi koneluettavia erikoisalojen ja yleisiä sanakirjoja.

Hakutilanteessa tapahtuvan hakukriteerin kääntämisen lisäksi monikielisyyttä voidaan tukea etukäteen, jos esimerkiksi on käytettävissä *automaattinen kielenkääntöjärjestelmä*. Kääntämällä (vaikka jonkin verran vajavaisesti) dokumentti hakijan äidinkielelle ja kohdistamalla hakukriteeri tähän (kenties puutteellisesti käännettyyn) dokumenttiin, voidaan kuitenkin haun tuloksena näyttää alkukielinen oikea dokumentti.

3.3 Dokumenttien automaattinen luokittelu

Toinen tiedonhaun kaltainen tehtävä on *dokumenttien automaattinen luokitus* sisältönsä perusteella. Tällaisesta on hyötyä esim. lehden toimituksessa, jonne saapuu jatkuvana virtana uutisviestejä. Yksittäiset viestit olisivat merkityksellisiä vain tietyille toimittajille ja kukin viesti pitäisi ohjata oikealle pöydälle käsiteltäviksi. Tietoja arkistoitaessa halutaan myös usein liittää niihin luokituksia, joiden arvellaan helpottavan niiden uudelleen löytymistä.

Dokumentin sisältöä voidaan kuvata sen tekstissä esiintyvien sananmuotojen (tai lekseemien) frekvenssien perusteella ottamatta huomioon saneiden keskinäistä järjestystä. Tiettyjen sanojen esiintyminen lisää todennäköisyyttä kuulua tiettyihin luokkiin.

Dokumentin luokittelu, niin kuin moni muukin tähän teemaan liittyvistä sovelluksista, on tehdään usein dokumenteissa esiintyvien sananmuotojen perusteella. Varsinkin englanninkielisille dokumenteille tämä toimiikin kohtuullisen hyvin, koska siinä kielessä sanat taipuvat kovin niukasti. Jos dokumentissa puhutaan tietystä henkilöstä tai asiasta, niin vastaavan sanan perusmuoto todennäköisesti esiintyy siinä. Toisaalta myös erilaisten sananmuotojen frekvenssit ovat isompia, kun kunkin lekseemin erilaisia sananmuotoja on vain vähän. Riittävän kokoiset frekvenssit ovat tärkeitä tilastollisille menetelmille.

Monissa muissa kielissä sanat taipuvat runsaammin, jolloin erilaisia sananmuotoja on väistämättä suhteessa enemmän ja yksittäisten sananmuotojen frekvenssit pienempiä. Luokittelemisen uhkana on se, että jos erilaisia

sananmuotoja on paljon, useimmat niistä esiintyvät vain yhdessä tai parissa dokumentissa. Kun saman lekseemin eri muodot ovat erillisiä yksiköitä, dokumentit näyttävät lukujen valossa erilaisemmilta, kuin ovatkaan.

Taipuvissa kielissä kannattanee käyttää sananmuotojen sijasta niiden perusmuotoja (eli lemmoja) yksikköinä. Erilaisia perusmuotoja on vähemmän kuin erilaisia sananmuotoja. Samaa asiaa käsittelevillä dokumenteilla on todennäköisemmin yhteisiä sanojen perusmuotoja kuin yhteisiä sananmuotoja. Tämä korostuu, jos dokumentit ovat lyhyehköjä.

Eräs useimpia ihmisiä koskettava luokittelun tarve liittyy ns. *roskapostiin*, jota tulvii enenevässä määrin sähköisiin postilaatikoihin, vaikka asianomaiset eivät suinkaan tällaisia viestejä haluaisi. Viestien karsinnassa voidaan käyttää monenlaisia keinoja. Useimmiten yhdistellään erilaisia tietoja, mm. viestin lähettäjän osoitetta, isojen kirjainten käyttöä, viestin koostumista pelkistä kuvista ja erityisiä sananmuotoja (kuten VIAGRA, NIGERIA, …). Roskapostin lähettäjät osaavat kuitenkin väärentää monia tietoja, esim. lähettäjän osoitteen. Sisällön perusteella tapahtuva karsinta voi toimia varsin tyydyttävästikin, sikäli kun ohjelmalla on ajan tasalla käytettävä malli tyypillisimmistä roskaposteista. Eräs tyypillinen roskapostin laji sisältää paljon keinotekoista sisältöä eli sananmuotoja, muttei mitään mielekästä viestiä. Roskapostin tunnistaminen on haaste kieliteknologialle, vaikka koko ongelmaa voidaan tietysti ratkoa muillakin teknologioilla.

Roskapostit ovat enimmäkseen englanninkielisiä, kieliteknologian saati monikielisen kieliteknologian tarve ei ole vielä herännyt roskapostien tunnistamiseksi. Roska-

postin tunnistamisen nykyinen teknologia koostuu joko yksittäisistä nokkelista oivalluksista tai tunnetuista tilastollisista tai todennäköisyyslaskentaan perustuvista malleista. Nokkeluudeksi voi laskea sen tapaiset menetelmät, joissa perustetaan sähköpostiosoitteita vain roskapostin vastaanottamiseksi ja laitetaan osoitteet verkkoon ja keskusteluryhmiin näkyville. Tällaisiin osoitteisiin tulevat viestit ovat siten roskaa ja ne osoitteet joista sitä tulee, voidaan laittaa mustalle listalle, jota postipalvelinten ylläpitäjät voivat käyttää erottamaan roskapostia. Vastaavanlainen menettely on ns. greylisting eli se, että kaikki uusista osoitteista tuleva posti käännytetään kerran takaisin. Oikea posti yleensä lähetetään uudestaan, roskapostia ei. Huomattakoon, että yllä kuvaillut menetelmät roskapostin tunnistamiseksi eivät ole varsinaisesti kieliteknologisia, vaan tavanomaiseen tietojenkäsittelyyn perustuvia.

3.4 Tekstin automaattinen tiivistäminen

Automaattinen tiivistelmän tekeminen dokumentista on useaan tarpeeseen hyödyllinen toimenpide. Pitkästä dokumentista haluttaisiin saada useinkin summittainen tieto siitä, mitä teksti pääasiassa käsittelee. Tiivistämisen pyrkimyksenä on, että *dokumentin tekstistä eristetään sellainen osa*, joka sopivien kriteerien mukaan sisältää dokumentin tyypillisintä tai tärkeintä asiaa. Tällaisen tärkeänä jakson tunnistamiseksi voidaan käyttää erilaisia hyvin yksinkertaisia kikkoja tai tilastollisia kriteereitä, kuten esimerkiksi sitä, että dokumentin ensimmäinen ja viimeinen kappale usein ovat tärkeitä (!), tai valitsemalla sellainen kappa-

le, jossa on runsaasti eri dokumentteja toisistaan erottavia sananmuotoja (tai lekseemejä).

Tekstejä erottelevina sananmuotoina tai hakusanoina voidaan pitää sellaisia, jotka eivät jakaudu eri dokumentteihin tasaisesti, vaan esiintyvät joissakin dokumenteissa yleisenä, mutta puuttuvat useista toisista kokonaan. Erottelevien sanojen pitää olla koko aineistossa riittävän yleisiä, jotta ne tehoaisivat moniin dokumentteihin (eikä vain yhteen). Toisaalta kaikkein yleisimmät sanat esiintyvät liian monissa dokumenteissa ollakseen hyödyllisiä. Dokumentista siis pitäisi etsiä sellaisia jaksoja, joissa dokumentille tyypilliset sanat esiintyvät. Tällaiset virkkeet tai kappaleet voivat olla lähtökohtana tiivistelmälle. Raakatiivistelmää voidaan edelleen työstää tunnistamalla sen lauserakennetta, jolloin voidaan valikoiden poistaa vähemmän keskeisiä aineksia sisältäviä osia.

Voisi hyvin ajatella, että nyt yleistyvissä edistyneemmissä matkapuhelimissa tiivistäminen saisi erittäin tärkeän sijan, koska laitteessa on pienehkö näyttö, jolla ei voi kohtuullisesti selata suurempaa tekstimäärää, eikä puhesynteesinkään avulla haluta kuunnella pitkiä jaksoja. Muutenkin tiedon hakemista voisi luontevasti täydentää tarjoamalla lyhyttä tiivistelmää, jonka perusteella jo useimmiten voisi päättää, haluaako vilkaista dokumenttia vai suoraan ohittaa sen.

3.5 Automaattinen hakemistojen muodostaminen

Dokumentin *indeksointi* eli *asiahakemiston laatiminen* on tavanomaisista kirjoista tuttu asia. Tekstistä pitäisi tunnis-

taa sisällön kannalta tärkeät termit ja erityisesti ottaa huomioon ne kohdat tekstistä, joissa termiin liittyvästä asiasta puhutaan intensiivisesti, esim. määritellään vastaava käsite. Tiedämme kokemuksesta, että kirjojen takana olevien hakemistojen tekeminen on ilmeisen vaikeaa, koskapa monissa kirjoissa nämä hakemistot ovat kovin puutteellisia. Automaattisilla menetelmillä voitaisiin ehkä päästä keskimääräistä käsintehtyä hakemistoa parempaankin laatuun.

Hakemistoon pitäisi toisaalta valita ns. *termejä* eli sellaisia (usein yhdestä tai parista sanasta koostuvia) ilmauksia, jotka sisältävät suhteellisen paljon informaatiota. Tätä voidaan arvioida yksinkertaisimmillaan vertaamalla termin esiintymien jakautumista kaikkien dokumenttien joukossa. Parempia ovat sellaiset ilmaukset, joiden esiintymät keskittyvät osaan dokumenteista kuin sellaiset, jotka jakautuvat jokseenkin tasaisesti kaikkiin dokumentteihin. Tällaisten tilastollisten kriteerien lisäksi hyödyllistä tietoa on osoitettu olevan myös ilmausten esiintymisellä tietynlaisissa lauseasemissa tai tiettyjen sanojen argumentteina tai ilmauksissa, jotka tyypillisesti ovat jonkun käsitteen määrittelyitä tms. Enemmän asiasta, ks. (Lahtinen, 2000).

3.6 Tiedon automaattinen eristäminen

Tiedon eristämisellä (engl. information extraction) tarkoitetaan sitä, että dokumenttien teksteistä eristetään tietyntyyppisiä ilmauksia kuten henkilöiden tai yritysten nimiä eli *nimettyjä kohteita* (engl. named entity) tai esim. tuotteiden julkistamisia tai henkilöiden nimittämisiä firmoissa uusiin tehtäviin. Tiedon eristämisessä aloitetaan

usein yksinkertaisista nimellä varustetuista kohteista kuten henkilöistä, firmoista tai tuotteista. Ongelmana näissä on yleensä se, että samasta kohteesta voidaan käyttää erilaisia ilmauksia, täydellisempiä ja lyhyempiä ja monella tavalla vaihtoehtoisia muotoja aina siihen, että PÄÄJOHTAJA THOMAS WATSON voidaan kirjoittaa lyhyemmin THOMAS WATSON tai PÄÄJOHTAJA tai HÄN. Näiden alla olevien kohteiden identifiointi on kuitenkin välttämätöntä, kun määritellään ja tunnistetaan isompia kokonaisuuksia vastaavia kohteita, kuten esim. IBM JULKISTI BLUETOOTH-TEKNOLOGIAN tai SONY JA ERICSSON FUUSIOITUIVAT.

Tiedon eristämisellä usein täsmennetään tiedon seulontaa silloin, kun esim. laajasta uutisvirrasta halutaan poimia esiin yrityksen oman toiminnan kannalta olennaisia uutisia (engl. business intelligence). Vastaavasti esimerkiksi rakennettaessa vaikkapa perinteistä lehtileikearkistoa vastaavaa tietojärjestelmää, voitaisiin hyödyntää tiedon eristämistä. Tiedon hakuun liittyvää valikointia voisi tarkentaa, jos voidaan viitata tiedon eristämisen tuloksena syntyviin relaatioihin. Tiedon eristämiseen liittyy usein *käsitehierarkioiden* laatimista ja hyödyntämistä. Käsitehierarkiat (engl. ontology) edesauttavat dokumenttien sisältämän tiedon tunnistamista ja rakenteistamista.

3.7 Hypertekstin ja semanttisen WEBin merkintöjen tuottaminen

Hypertekstiä muodostettaessa pitäisi tunnistaa, missä on *tietoa, johon kannattaa viitata ja mistä tällaisia viitteitä olisi tarpeen tehdä*. Ne, jotka sattuvat tuntemaan *verkkosivujen merkkauksessa* (engl. markup) käytettyä HTML-kieltä, tunnistavat että edellisiä viittauksia merkataan A-alkioiden NAME-attribuutilla ja jälkimmäisiä HREF-attribuutilla.

Semanttiseksi WEBiksi (engl. semantic web) kutsutaan XML-kielen puitteissa tapahtuvaa verkkosivujen merkkausta, jossa osoitetaan yleisellä tavalla erilaisia *suhteita* sivujen ja kohteiden välillä. Standardeja laaditaan siitä, millaisia merkintöjä verkkosivuihin tehtäisiin. Merkkaukset perustuvat yleensä käsitehierarkioihin (joita myös ontologioiksi kutsutaan). World Wide Web Consortium, W3C[2] on määritellyt merkkaustavat (mm. RDF eli *Resource Description Framework*), joilla semanttisen webin relaatioita ilmaistaan. Olennaista on se, että tavanomaisen linkin sijasta viittauksessa ilmaistaan relaation laji. Erilajiset relaatiot puolestaan kuvataan omilla formalismeillaan (esim. OWL eli *Web Ontology Language*), joka määrittää näiden relaatioiden suhteita toisiinsa.

Semanttinen web on näin määriteltynä lähinnä keino *ilmaista* näitä suhteita ja merkityksiä. W3C:n määrittelyt eivät varsinaisesti kerro, miten tällaisia merkkauksia voitaisiin *tuottaa* olemassa oleviin teksteihin, eikä myöskään

[2]ks. `http://www.w3.org/2001/sw/`

auta siinä, miten suurten tietomassojen tällaisia relaatioita merkkausta pitäisi *hyödyntää*, jos siis olisi saatavilla.

Kyseiset kaksi aluetta jäänevät kieliteknologian haasteiksi. Tiedon eristämisen keinoilla ja jäsennysmenetelmillä voitaisiin tällaisten merkkausten tekemistä avustaa tai osin automatisoida. Vastaavasti luonnollisen kielen *keskustelujärjestelmä* voisi olla oikea lähestymistapa saada semanttisen webin muodossa olevasta tiedosta irti vastauksia hakijan kysymyksiin. Näin siksi, että kysyjä tuskin pystyy kerralla ilmaisemaan riittävän täsmällistä kysymystä, mutta sopivan palautteen ja täsmentävien kysymysten jälkeen ehkä kylläkin.

3.8 Kieliteknologiset menetelmät, joita tarvitaan tekstitiedon hallinnassa

Yhteisenä tekijänä tämän luvun menetelmille on se, että ne perustuvat suoraan sananmuotojen esiintymiseen ja yleisyyteen. Niitä on sovellettu ensiksi englanninkielisiin dokumentteihin ja varsin pitkään jokseenkin ilman mitään kieliteknologisia apuvälineitä. Tämä onkin ymmärrettävää, koska englannin kielessä sanat taipuvat niukasti, minkä vuoksi aineiston kaikkien erilaisten sanamuotojen määrä on kohtuullinen ja erityisesti yhdestä hakusanasta on vain pieni määrä eri sananmuotoja.

Varsinaisten kieliteknologisten morfologisten jäsentimien tms. sijasta onkin varsinkin alkuaikoina sovellettu, esimerkiksi *sananmuotojen typistystä* (engl. stemming) tai muuta melko karkeata käsittelyä. Kirjallisuudessa tunne-

tuin näistä lienee ns. Porter stemmer.[3]

Yleinen tarkastelutapa on se, että dokumentteja kuvataan niiden juoksevan tekstin sijaan matemaattisesti *vektoreilla*, joiden alkiot kuvaavat kunkin mahdollisen sananmuodon esiintymiskertoja koko dokumentissa. Vektori on matematiikan käsite, jolla on suunta ja pituus. Vektoreita esitetään käytännössä komponenttiensa avulla, ne ovat siis vakiomittaisia lukusarjoja, joissa tietty paikka sarjassa vastaa tiettyä ominaisuutta (tai ulottuvuutta). Sananmuotoihin perustuvat vektorit ovat kovin pitkiä, eli sisältävät kymmeniä tuhansia alkioita (yhden alkion kutakin esitysmuotoon valittua sananmuotoa kohden). Vektoreissa ei ole tietoa sanajärjestyksestä, mutta toisaalta ne soveltuvat helposti monenlaisiin tilastollisiin menetelmiin. Ajatuksena on mm. se, että tällaisina vektoreina tekstejä voidaan helposti verrata toisiinsa. Tekstit voivat nimittäin olla enemmän tai vähemmän samansuuntaisia keskenään juuri niin, kuin vektoritkin voivat olla.

3.8.1 Sanojen taipumisen ja yhdyssanojen vaikutus tiedonhakuun

Monissa kielissä sananmuotoja on lukuisia kohti paljon enemmän kuin englannissa. Aiemmin todettiin suomen kielen erilaisten sananmuotojen huikea määrä. Vaarana pelkkien sananmuotojen käyttämisessä taipuvissa kielissä on se, että vaikka kaksi dokumenttia puhuisi samasta asiasta ja niissä käytettäisiin samoja hakusanoja asian ilmaisemiseen, hakusanat voivat helpostikin olla eri taivutusmuo-

[3] Ks. Wikipedia tai http://tartarus.org/~martin/PorterStemmer/

dossa. Tilastot, joita esiintymistä lasketaan voivat hajota sananmuodoilla niin moneen eri yksikköön, että useimmat niistä ovat vain nollan tai yhden esiintymän suuruisia (ja sellaisina moniin tarkoituksiin hyödyttömiä).

Taipuvassa kielessä, kuten suomessa, päällimmäisin vaara on se, että hakujen *saanti* sananmuotoja käytettäessä on heikko. Jos taipumista pyritään neutraloimaan sillä, että haetaan kokonaisen sananmuodon sijasta prefiksillä, paranee saanti, mutta varsinkin lyhyillä sanoilla *tarkkuus* kärsii. Esim. suomen sana VESI ei toisi hakuterminä sen muita taivutusmuotoja kuten VETTÄ tai VEDESSÄ. Jos sanaa pitää typistää riittävästi, jotta kaikki sen taivutusmuodot tulevat mukaan, jää jäljelle vain VE-, joka prefiksi puolestaan on aivan liian yleinen ja tuottaa enimmäkseen muita sananmuotoja kuin haluttuja.

Suomessa ja monessa muussakin kielessä yhdyssanat kirjoitetaan yhteen ja yhdyssanoja on kahdenlaisia: (1) *vakiintuneita*, joita käytettäisiin mieluusti kokonaisina hakutermeissä (esim. PENKINPAINAJAISET) ja (2) spontaanisti muodostettuja, usein aivan *kertakäyttöisiä* (esim. PLASMARIITA), joissa jälkiosa olisi mieluusti tunnistettava ja kelpuutettava hakutulokseen (kyse on RIIDASTA ja PLASMASTA eli televisioiden plasmanäytöistä). Yhteen kirjoitettujen yhdyssanojen takia hakujen saanti on vaarassa pienentyä. Useimmat tiedonhakujärjestelmät eivät edes salli hakutermin katkaisua alusta käsin, mikä voisi auttaa löytämään yhdyssanojen jälkiosia. (Eikä tämä taipuvissa kielissä yleensä ratkaisisi ongelmaa.)

3.8.2 Morfologisen jäsentimen käyttäminen

On luontevaa, että taipuvissa kielissä sananmuotojen sijasta käytettäisiin lekseemejä eli hakusanoja, jotteivät lukumääristä muodostetut vektorit kasvaisi tähtitieteellisen pitkiksi ja sisältäisi enimmäkseen nollia. Hakusanoihin siirtyminen on mahdollista morfologisen jäsentimen avulla. Tällainen jäsennin (tai analysaattori) ottaa syötteekseen tietyn kielen sananmuotoja ja päättelee niistä sanakirjan ja taivutussääntöjen perusteella, mikä sananmuodon perusmuoto, ts. alla oleva lekseemi voisi olla. Perusmuotoihin palautettuna tilastot saavat enemmän profiilia ja sokean tilastolliset menetelmätkin toimivat paremmin. Morfologinen jäsennin tunnistaa myös yhdyssanojen osat, jolloin niitäkin voidaan käyttää haun kriteereinä.

Haun *saantia* saadaan *parannetuksi* morfologisella jäsentimellä, kun lekseemin kaikkien taivutusmuotojen esiintymät löydetään. Sikäli, kun sananmuodot ovat moniselitteisiä, pelkän morfologisen jäsentimen käyttö voi toisaalta jossakin määrin *heikentää tarkkuutta*, sillä tekstin sane tulee indeksoiduksi kaikkien tulkintojensa mukaisesti. Siten löytäisimme dokumentteja, joissa puhutaan höyrymoottoreista (KONEISTA), kun haemme KONI-sanan esiintymiä.

Toinen ongelma liittyy siihen, että morfologinen jäsennin ei välttämättä tunne kaikkia uudissanoja ja nimiä eli se ei anna mitään tulkintaa tällaisille sananmuodoille. Morfologisen jäsentimen rinnalla voikin olla ns. *morfologinen arvain*, joka päättelee, millainen lekseemi voisi periaatteessa tuottaa havaitun sananmuodon taivutusmuotonaan.

Tällaisten keinojen avulla kieliteknologia neutraloi luonnollisen kielen tuottamaa vaihtelua ja palauttaa taipu-

van kielen hakutehtävän samanlaiseksi kuin se vähemmän taipuvassa kielessä on.

3.8.3 Hakuvartaloiden muodostaminen

Eräs vaihtoehtoinen, morfologista jäsennintä yksinkertaisempi menetelmä, jota käytetään taipuvien kielten kohdalla, on ns. *hakuvartaloiden muodostaminen*. Useat tiedonhakujärjestelmät (joskaan eivät läheskään kaikki) osaavat hakea paitsi täsmällisten sananmuotojen perusteella, myös prefiksien avulla. Tällöin kelpuutetaan sananmuotoja, jotka alkavat annetulla prefiksillä. Joissakin tapauksissa taipumisen voi neutraloida katkaisemalla sanan vartalon sopivasta kohdasta niin, että tällainen tyvi pyydystää sanan kaikki taivutusmuodot. Tavallinen käyttäjä ei kuitenkaan ole kovin luotettava tällaisten katkaisujen tuottamisessa. Hakuvartaloita tuottava ohjelma tekee tämän käyttäjän puolesta. Esimerkiksi, kun käyttäjä syöttää substantiivin KAUPPA, hakuvartaloita muodostava ohjelma tuottaa seuraavat prefiksit:

KAUPPA-
KAUPA-
KAUPPO-
KAUPO-

Haku suoritetaan sitten näiden prefiksien avulla.

Hakuvartaloiden ja morfologisen jäsentimen välimaastoon on myös ehdotettu menetelmiä, joissa säännöllisen lausekkeen avulla rajataan kohteena olevia sananmuotoja tarkemmin kuin hakuvartaloilla, muttei yhtä tarkkaan kuin, mitä morfologinen jäsennin tekisi.

Toisaalta on tuotannossa olevia tiedonhaun järjestelmiä, joissa yhdistetään sekä hakuvartalomekanismi että morfologinen jäsennys ja vielä näiden lisäksi hakutulosten suodatus sen perusteella, että hakuvartalon perusteella löydetty sananmuoto voisi todella olla haettavan sanan jokin taivutusmuoto.

3.8.4 Lauseyhteyksien hyödyntäminen

Yllä olevat menetelmät paransivat ennen kaikkea *saantia*, joka on vaarassa jäädä huonoksi taipumisen takia. Kieliteknologiaa voidaan käyttää myös *tarkkuuden* parantamiseen. Tilastollisia menetelmiä käytettäessä saadaan oikeammin osuvia lähtötietoja, jos tulkintoja voidaan riittävän luotettavasti yksiselitteistää tai jos sanojen käytöstä saadaan lauseenjäsennyksen avulla yksityiskohtaisempaakin tietoa. Esimerkiksi raaka tilastomenetelmä tyytyisi laskemaan substantiivin kertoja, kun kieliteknologian kanssa tähän voitaisiin liittää esim. tieto siitä, että sana esiintyy lauseessaan esim. *objektina tai aivan tietynlaisen verbin subjektina*. Tällä tavalla kieliteknologia voi antaa tilastomenetelmille *paremmin erottelevia* lähtömuuttujia. Näitä menetelmiä tarvitaan erityisesti, jos pyritään tuottamaan puoliautomaattisesti tai automaattisesti semanttisen webin merkkauksia tai kirjojen tai muiden laajojen dokumenttien hakemistoja.

3.9 Tulevaisuudennäkymiä

Tiedonhaku sinänsä on vuosikymmenien ikäistä ja lähtenyt liikkeelle kauan ennen Internetin syntyä. Tietynlaisten

tietojen etsiminen ja löytyminen onnistuu melkeinpä liiankin hyvin, esimerkiksi useiden nimien tai vakiintuneiden akronyymien. Taivutuksen neutralointi on myös ollut käytössä jo pian parikymmentä vuotta eräissä suomenkielisissä tiedonhakujärjestelmissä.

Myös jotkut tavallisiin sanoihin perustuvat haut onnistuvat erinomaisesti, kunhan sananmuotoja on riittävän monta. Esimerkiksi Googlella etsimällä löytyi helposti esiintymiä englanninkielisestä vitsistä, joka kuuluu suunnilleen seuraavasti:

> Three old men were talking:
> A: ITS WINDY TODAY
> B: NO, I THINK IT'S THURSDAY
> C: SO DO I, LET'S GO AND HAVE A BEER
> And so they went to a bar.

Kaikki sanat ovat perin tavallisia, mutta niiden yhdistelmä ei enää olekaan ja hakukone tuottaa melko hyvin juuri tämän vitsin sisältäviä sivuja. Jos haussa vielä kytkee tiettyjä peräkkäisiä sanoja lainausmerkeillä yhteen, saa melko lailla puhtaasti vain tuon vitsin esiintymiä.

Silti, vaikka osa hauista onnistuu hyvin, kaikki eivät suinkaan. Esimerkiksi Googlen käyttö puheliittymän kautta olisi äärimmäisen hidasta ja piinallista, koska hakukoneen käyttöliittymä hyödyntää suurta kuvaruutua ja sitä, että käyttäjä voi silmäilemällä helposti tehdä ratkaisuita siitä, mitkä haun tuottamista sivuista ovat ylipäätänsä mahdollisia.

Puhekäyttöliittymään tarvittaisiin toisaalta parempaan tarkkuutta, koska selaaminen on hidasta ja toisaalta interaktiivista dialogia, koska käyttäjän on vaikeaa kerralla an-

taa riittävän rajaavaa hakukriteeriä. Ehkä tämä tulee mahdolliseksi semanttisen Webin puitteissa siten, että toisaalta merkkauksia voidaan saada riittävästi verkkoon kielikeknologisilla työkaluilla ja toisaalta tällaisiin merkkauksiin voitaisiin kohdentaa tarkempia hakuja, joita vuorovaikutteisella järjestelmällä hyödynnettäisiin.

Luku 4

Puheteknologia ja kieliteknologia

4.1 Puheen olemuksesta

Puhuttu kieli on ensisijaisempaa kuin kirjoitettu, sillä ihmiset oppivat vuoden tai parin ikäisinä puhumaan, mutta kirjoittamaan paljon myöhemmin, jos ollenkaan. Suurimmalla osalla maailman kielistä ei ole lainkaan kirjakieltä eikä kirjoitettua muotoa. Vaikka kirjakieli olisi olemassa, kaikissa kulttuureissa ei kirjoituksella silti ole tärkeää sijaa. Voitaisiin jopa spekuloida, että nykyisessä länsimaisessa teknologisessa kulttuurissa kännyköiden, kuvan ja puheen siirtämisen ja taltioimisen helpottuessa kirjoitetun kielen merkitys taas vähentyisi.

Puhe on yleensä vähemmän suunniteltua ja *kertakäyttöisempää* kuin kirjoitettu kieli. *Puhe häipyy siinä missä kirjoitus säilyy*. Pääsääntöisesti kaikkia kieliä puhutaan tai niitä ainakin niitä on puhuttu. Poikkeuksena ovat ehkä

kuurojen viittomakielet, joita niitäkään ei yleensä kirjoiteta, mutta ei myöskään puhuta.

Puhesignaalin tunnistaminen — eli oikein kuuleminen ja kuullun ymmärtäminen — on *ihmiselle helppoa ja luontaista*. Oikein kuuleminen osoittautuu kuitenkin hankalaksi toteuttaa tietokoneella. Parhaatkaan signaalinkäsittelymenetelmät eivät erota yksittäisiä äänteitä (eivätkä myöskään kokonaisia sanoja) toisistaan yhtä hyvin kuin ihminen. Kone ei yllä samaan tarkkuuteen kuin ihmiskorva, vaikka signaalia käsiteltäisiin kuinka perusteellisesti tai isoilla ja nopeilla koneilla hyvänsä. Vastaavasti, tietokonetta on vaikea opettaa puhumaan niin ihmisenkaltaisesti, ettei konemaisuutta huomaisi.

Kuitenkin, rajoituksistaan huolimatta automaattinen puheentunnistus ja puhesynteesi ovat kehittyneet sekä mahdolliseksi että hyödylliseksi ja molemmilla on jo paljon toimivia sovelluksia. Tunnistuksen tarkkuus ja synteettisen puheen luonnollisuus sekä paranevat jatkuvasti, vaikka täydellisyyteen on paljon matkaa.

4.1.1 Kieltä puhutaan eri tavalla kuin kirjoitetaan

Se kieli, jota normaalisti puhutaan, on muodoltaan ja rakenteeltaan erilaista kuin kirjoitettu kieli mm. siinä, että:

- puheenvuorot ovat lyhyehköjä tai aivan lyhyitä ja tiukasti *puhetilanteeseen sidottuja*, usein syntaktisesti *epätäydellisiä* tai epäkieliopillisiakin (koska suunnitelmaa muutetaan kesken kaiken),

- *sanasto* on erilaista, usein myös yksinkertaisempaa,

- sanojen *taivutus* voi olla erilaista,
- *murre-eroja* esiintyy puheessa herkemmin enemmän kuin kirjoitetussa tekstissä ja
- puhujien väliset *ääntämyksen* erot ovat suurempia kuin kirjoitetun tekstin erot, mutta *myös saman puhujan* eri tuotosten välillä on eroja runsaasti.

Puhekieliä on tutkittu ja kuvattu *vähemmän* kuin kirjoitettuja, eikä puhekielistä ole vastaavia kielioppeja kuin kirjoitetusta. Sen vuoksi puhekielten ominaisuuksia tunnetaan puutteellisesti ja tällä alueella tarvitaan paljon työtä.

4.1.2 Puhe fysikaalisena signaalina

Puhe on ääntä ja se syntyy ihmisen puhe-elimissä suunnilleen siten, että *kurkunpäässä* syntyvää pärinä muokkautuu eri äänteiksi sen mukaan, millaisessa asennossa varsinainen *ääntöväylä* kulloinkin on, ks. teoksessa Jurafsky & Martin (2008) kuva 7.3 sivulla 219. *Kielen* eri asennot muuttavat ääntöväylää siten, että erilaiset *taajuudet* vahvistuvat ja alun perin samanlaisesta kurkunpään äänestä muotoutuu tunnistettavia eri äänteitä. Kielen asentojen lisäksi reitti *nenään* voidaan avata ja sulkea ja huulilla tai kielellä voidaan ilman kulkua *estää* hetkeksi kokonaan tai lähes kokonaan. Tätä kaikkea kuvaillaan tarkemmin *fonetiikan* kursseilla ja sen alan oppikirjoissa. Myös Wikipedian suomen- ja erityisesti englanninkielisistä fonetiikkaa koskevista artikkeleista löytyy hyviä havainnollistuksia ja mm. eri kielissä esiintyvien vokaalien ääninäytteitä.[1]

[1] Ks. esim.: `http://en.wikipedia.org/wiki/Vowel`

Äänen tallentaminen ja toistaminen

Ääni on fysikaalisesti katsottuna vain *ilmanpaineen nopeita vaihteluita*. Äänisignaalia voidaan kuvata *käyrällä*, joka noudattelee näitä paineen vaihteluita, esim. teoksessa Jurafsky & Martin (2008) pientä vokaalin viipaletta kuvaava kuva 7.13 sivulla 232 ja kokonaista virkettä tai sanaa kuvaavia käyriä kuvissa 7.17 s. 236 ja 7.18 s. 237.

Kun puheen ääni on syntynyt, se on siis ilmanpaineen nopeista vaihteluista koostuva signaali, ja siihen sisältyy *kaikki tarvittava tieto äänen toistamiseksi* samanlaisena. Siten mikrofoni voi muuttaa paineen värähtelyt sähkön värähtelyiksi, jotka voidaan *tallentaa* joko koodattuna lukusarjana tietokoneen muistiin tai CD-levylle tai magneettisuuden vaihteluina ääninauhalle tai vaikka lähettää radioaaltoina eetteriin. Eri muodoissa oleva signaali saadaan vahvistimella ja muilla äänentoistolaitteilla korvin kuultavaksi ääneksi kaiuttimesta tai kuulokkeista siten, että se *kuulostaa* hyvinkin *samanlaiselta puheelta* kuin, mitä se alunperin tallennettaessa oli.

Kaikki fysikaalinen tieto puheäänestä on sen signaalissa eli käyrässä, joka voidaan kuvata tarkasti *sarjana lukuja*, jotka kuvaavat sitä, kuinka korkealla tai matalalla signaalia kuvaava käyrä kunakin hetkenä on. Lukuja eli näytteitä tarvitaan yleensä 8 000 – 50 000 sekunnissa sen mukaan, kuinka laadukkaana ääni halutaan tallentaa. Tässä lukusarjassa on valtavasti *enemmän tietoa*, kuin mitä tarvitaan äänteiden ilmaisemiseen, eli ehkä tuhansia lukuja, kun yksi ainoakin niistä riittäisi osoittamaan kulloisenkin äänteen. Vaikka signaalissa on äännettä kohti tuhansia lukuja, lukusarjasta ei silti ilmene millään *helpolla* tavalla se, mikä äänne kussakin kohdassa on meneillään.

Ihmisen puheääni ei kuitenkaan ole mitä tahansa ääntä. Sillä on ominaispiirteitä, joita käytetään hyväksi mm. ääntä koodattaessa matkapuhelimissa. Kuulemisen kannalta olennaisia vaihteluita voidaan eristää erilaisilla menetelmillä, jotka hyödyntävät puheäänen erityislaatua tallentamalla värähtelystä tietoa sen toistumisesta erityisesti värähtelyjen taajuuksista. Jotakin tällaista ihmiskorvakin tekee, vaikkakin ilmeisesti omalla tavallaan.

Tallennettavissa oleva signaali sisältää kaiken tarvittavan tiedon puheen toistamiseksi varsin luonnonmukaisena. Puhe näyttäisi siis fysikaalisesti katsottuna olevan varsin ongelmaton ja helposti hallittava ilmiö. Aivan näin yksinkertaista puhe ei kuitenkaan ole, vaan kuulijalle ilmeinen ja tärkeä sisältö on koodattuna kovin monimutkaisesti tähän signaaliin. Jostakin syystä (ainakin lukutaitoinen) ihminen hahmottaa puheesta samana pitämiään äänteitä esim. vokaalin A *samana* foneemina, vaikka sen toisinnot voivat olla fysikaalisesti hyvinkin *paljon toisistaan poikkeavia* esiintyessään eri yhteyksissä. Ääritapauksessa ihminen voi mielestään kuulla puheessa äänteitä, joita siinä todellisuudessa ei ole, koska hän ikään kuin täydentää kuulemaansa ja muilta osin tunnistamaansa sen perusteella, mitä muuten tietää.

Äänisignaalin muokkaaminen, spektri ja spektrogrammi

Puhesignaalia voidaan paitsi välittää ja tallettaa sähköisesti sekä toistaa, myös tutkia muokata varsin monipuolisesti. Paremman äänen laadun aikaansaamiseksi ohjataan tavanomaisissa kotistereoiden kaiuttimissa äänen korkeimmat taajuudet jakosuodattimella pieneen ja matalammat taa-

juudet suurempaan kaiuttimeen.

Signaalinkäsittelyä varten on olemassa runsaasti matemaattisia menetelmiä, esimerkiksi kompleksilukuihin perustuvia erilaisia muunnoksia ja operaatioita, mm. *Fourier-muunnos*. Tällaisten menetelmien avulla voidaan alkuperäisestä signaalista saada johdetuksi muita esitysmuotoja, jotka tuovat esille äänteiden tärkeitä ominaisuuksia helpommin tunnistettavassa muodossa. Joistakin tällaisista esitysmuodoista voidaan tarvittaessa palauttaa alkuperäinen äänisignaali takaisin, eli tällaisilla muunnoksilla on käänteismuunnos. Ihmiskorvan kannalta tällainen ei aina ole välttämätöntä, sillä korva ei kiinnitä huomiota esimerkiksi signaalin vaiheisiin (jotka riippuvat myös siitä, kuinka kaukana korva on äänilähteestä).

Äänisignaalin jakaminen eri taajuuksiksi voidaan tehdä hienojakoisemminkin kuin vain jakosuodattimella kahteen erikokoiseen kaiuttimeen. Tyypillisesti taajuuksien erottelu tehdään juuri em. Fourier-muunnoksella, joka jakaa äänen eri taajuuksiinsa. Muunnos kohdistetaan useimmiten signaalin pieneen määrämittaiseen siivuun kerrallaan, josta lasketaan *spektri* eli hajotetaan ääninäyte eri taajuuksiinsa, ks. Jurafsky & Martin (2008), missä kuvassa 7.22 sivulla 238 on englannin [ae]-vokaalista sen eri taajuuksien jakautuma ja kuvassa 7.25 sivulla 240 vastaava [ij]-vokaalista.

Kokoamalla puhutun sanan tai lauseen eri viipaleista lasketut spektrit nauhaksi saadaan ns. *spektrogrammi*, ks. Jurafsky & Martin (2008) kuvaa 7.23 sivulla 239, jossa kolmen eri vokaalin spektrogrammit sekä kuvaa 7.24 sivulla 240, jossa kokonaisen virkkeen spektrogrammi. Spektrogrammissa voidaan visualisoida puhetta siten, että

puheen edetessä vaaka-akselia pitkin kunkin hetken taajuusjakautuma eli spektri on pystyakselilla esitettynä tummuusasteikolla. Tummat alueet spektrissä merkitsevät sitä, että niitä taajuuksia on runsaasti.[2] Varsinkin vokaaleilla on selkeät ns. *formantit* eli taajuudet, jolle äänen energia keskittyy. Äänihuulten omaa taajuutta kutsutaan formantiksi F0 ja ensimmäistä sitä korkeammilla taajuuksilla olevaa keskittymää F1:ksi jne. Yleensä vain ensimmäiset näistä formanteista nimittäin F1, F2 ja toisinaan F3 ovat äänteiden tunnistamisen ja luokittelun kannalta mielenkiintoisia.

Äänisignaalia voidaan paitsi tarkastella spektrinä, myös esim. muunnella vahvistamalla tai heikentämällä tiettyjä taajuuksia siihen tapaan, kuin edellä todettiin ääntöväylässä tapahtuvan silloin, kun ihminen puhuu. Ihmispuheen kaltaisen puheen tuottaminen on kuitenkin varsin vaativa tehtävä, kuten tuonnempana saamme nähdä.

4.1.3 Puheessa esiintyvä vaihtelu

Puhe on vaihtelevaa toisaalta tahattomasti ja toisaalta tarkoituksellisesti. Eri ihmiset ääntävät puhetta vähän eri tavalla jo senkin takia, että heidän ääntöelimensä ovat kooltaan tai muodoltaan vähän erilaisia. Sen lisäksi kukin on oppinut puhumaan vähän omalla ominaisella tavallaan.

Miesäänet ja naisäänet ovat erilaisia mm. siksi, että kurkunpään rakenteessa eroja, jotka vaikuttavat äänisignaaliin. Kurkunpäässä syntyvä taajuus on naisäänessä lähempänä formanttitaajuuksia, jonka takia naisäänen

[2] Vastaavia kuvia löytyy myös Wikipedian englanninkielisistä artikkeleista Spectrogram ja Frequency spectrum, ks. http://en.wikipedia.org/wiki/Spectrogram

spektri ei ole fysikaalisesti tarkasteltuna niin selkeä kuin miesäänen, vaikka ihmiskorva kuulee naisäänet vähintään yhtä selvänä kuin miesäänet.

Sama ihminen ääntää *eri kerroilla* saman ilmauksen eri tavalla, vaikka kuinka yrittäisi toistaa sen aivan samanlaisena. Äänteiden kestot, voimakkuudet, kielen ja huulten asennot ym. eivät tietenkään voi eri kerroilla olla täsmälleen samanlaisia.

Puhutussa kielessä käytetään myös herkemmin paikallisia ja *murteellisia* ääntämyksiä ja muotoja kuin kirjoitetussa. Puhelimen kautta tarjotaan nykyään palveluja keskitetysti koko valtakuntaa varten, jolloin kaikenlaisia paikallisia puhetapoja tulee vuoronperään, mistä sekä inhimillisen että tietokoneistetun palvelun täytyy selvitä.

Kuvittelemme yleensä, että puheessakin olisi jokseenkin *vakinaisia äänteitä* siten kuin kirjoitetussa tekstissä on kirjaimia. Vaihtelu on kuitenkin paljon suurempaa, kuin luulisi. Yksittäinen vokaali on aika erilainen sen mukaan, *mitä konsonanttia se seuraa tai edeltää*. Painollinen vokaali on samassa ympäristössäkin erilainen kuin painoton jne. Eri puhujien erilaisissa ympäristöissä tuottamat samoina pidettävät äänteet hajoavat siinä määrin laajalle, että eri äänteiden fysikaalisten ominaisuuksien perusteella laaditut jakaumat menevät osittain päällekkäin.

Vokaaleja luonnehditaan näissä yhteyksissä niiden taajuusspektrin mukaan tarkastelemalla kahta ensimmäistä ns. formanttia F1 ja F2, eli kahta alinta taajuutta, joilla on runsaasti energiaa. Vokaalit tunnistetaan fysikaalisesti yleensä näiden *formanttien avulla* ja tätä tarkoitusta varten piirretään kaksiulotteisia diagrammeja, joissa toisella akselilla on F1:n taajuus ja toisella F2:n taajuus. Vokaalit

ryhmittyvät tietyille alueille, mutta alueet eivät ole toisistaan irrallisia.

Kirjallisuudesta löytyy havainnollistuksia äänteiden luokittelusta ja päällekkäisyydestä, esim. artikkelissa Hillenbrand et al (1995)[3], teoksessa Harrington & Cassidy (1997) kuva 4.2 sivulla 70, kuva 4.9 sivulla 79 jne. sekä teoksessa Rabiner & Juang (1993) kuva 2.16 sivulla 27.

4.1.4 Prosodia

Puhe on muutakin kuin sarja peräkkäisiä äänteitä. Puheeseen liittyy tauotusta, eli siellä täällä sanojen välissä on *taukoja*. Useimmiten sanat kuitenkin äännetään yhteen ilman taukoa, vaikka kirjoituksessa sanojen välit ilmaistaan selvästi. Sanojen sisällä olevat klusiilikonsonanteista johtuvat tauot voivat hyvinkin olla pitempiä kuin sanojen väliin sattuvat nimenomaiset tauot, mikä paljastuu puhesignaalia mittaamalla. Oma intuitiivinen käsityksemme ei vastaa tätä fysikaalisesti mitattavaa todellisuutta, vaan sanojen sisäiset tauot saattavat tuntua lyhyemmiltä kuin ovat (jos ne ylipäätänsä mielletään tauoiksi).

Tiettyjä kohtia ilmauksista *painotetaan* osoittaen niiden tärkeyttä, mutta painotuksella osoitetaan myös epäsuorasti, miten ilmaus jakautuu sanoiksi. Kuulijasta ja puhujasta voi tuntua siltä, että painotus olisi ennen kaikkea *runsaampaa energiaa* puhesignaalin painotetussa kohdassa. Aina ei ole kysymys pelkästään tästä, vaan painotus voi ilmetä esimerkiksi myös äänteiden oletuksesta poikkeava-

[3] Kyseisen artikkelin *Acoustic characteristics of American English vowels* PDF-versio löytyy verkosta esim. artikkelin erään kirjoittajan kotisivulta

na kestona.

Kurkunpäästä johtuva Puheen ns. perustaajuus (F_0) vaihtelee ilmauksen kuluessa ja havaitaan puheen *sävelkulkuna*. Sen avulla merkitään ilmauksen loppumista (monissa kielissä laskevalla sävelkululla) ja ilmauksen sisällä olevaa rakennetta. Esimerkiksi ilmauksen yleisessä laskevassa sävelkulussa hypätään tarvittaessa ylöspäin sopivien kokonaisuuksien rajoilla. Sävelkorkeuden hidas vaipuminen ja sen nousut auttanevat kuulijaa hahmottamaan ilmausta oikein.

Yllä kuvattuja ilmiöitä kutsutaan yhteisesti *prosodiaksi*. Prosodisilla keinoilla puhetta ikään kuin *moduloidaan*, jotta kuulijan olisi helpompi hahmottaa sitä. Toisaalta prosodia itsessään aiheuttaa vielä lisää vaihtelua äänteiden fysikaalisiin ominaisuuksiin, kun äänteistä esiintyy erikestoisia, eri perustaajuudella olevia sekä eri lailla painotettuja versioita.

Eri kielissä ilmaistaan painotuksia eri keinoin ja sävelkulut voivat olla erilaisia. Yksittäisten kielten kuten suomen prosodiassa on paljon tutkittavaa ennen, kuin ymmärretään, millaista prosodiaa minkäkinlaiseen tilanteeseen ja ilmaukseen pitäisi soveltaa.

4.2 Puhesynteesi

Puhesynteesillä tuotetaan tietokoneen avulla ihmispuheen kaltaista ääntä. Ei ole itsestään selvää, millaista onnistuneen puhesynteesin pitäisi olla, sillä hyväntuntuiset päämäärät voivat olla ristiriitaisiakin.

- Taitavan *inhimillisen lukijan kaltaisuus*: Valistunut ihminen lukee tekstistä usein sellaista, mitä ei teks-

tiin ole eksplisiittisesti merkitty, esim. lukee lyhenteitä niiden täydellisessä muodossaan sekä esim. rahasummien, päivämäärien ja puhelinnumeroiden numerot kunkin omalla oikealla tavallaan.

- *Ilmeikkyys*: Konemainen puhe koetaan monotoniseksi. Toisaalta ilmeikkyys saattaa joskus ärsyttää tai häiritä kuulijaa, eikä sitä pidä liioitella.

- Selkeys eli se, että *kuulija kuulee viestin oikein*: Esimerkiksi matkapuhelimella kommunikoitaessa on usein hälyä häiritsemässä ja äänen laatu ei ole optimaalinen. Silloin varmaan selkeys on ansio ja lisää viestin mahdollisuuksia mennä oikein perille. Liioiteltu selkeys voi sekin olla häiritsevää, jos kuuluvuus on hyvää ja häiriötöntä.

- *Viritettävyys* ja muunneltavuus: Toiset käyttäjät voivat seurata nopeampaa synteesiä kuin toiset. Jotkut tekstit voivat olla helpompia seurata kuin toiset. Esimerkiksi näkövammaiset haluavat yleensä niin nopeata lukemista, kuin missä vielä pysyvät perässä — ja tämä voi olla paljon nopeampaa kuin ihmisen normaali puhe.

- *Ohjailtavuus* lukutilanteessa: Tekstiä puheeksi syntetisoitaessa tulee välillä vaikeampia kohtia, joita pitäisi ottaa hitaammin uudestaan. Tarvittaessa olisi voitava lukea tavaamalla tai kirjain kirjaimelta, jotta saadaan selvää, mitä hankalissa paikoissa oikein lukee.

4.2.1 Puhesynteesin sovellukset

Puhesynteesiä tarvitaan eriasteisena silloin, kun automatisoidaan laitteita tai palveluita. Synteesiin tarvittava tai sovellettu tekniikka voi vaihdella melkoisesti aivan yksinkertaisesta hyvinkin vaativaan, mutta tekniikoita käsitellään tuonnempana.

Kuulutukset ja tiedotukset yleisölle

”Neiti Aika” eli puhelimella saatava aikatiedotus hoidettiin aluksi ihmisvoimin, mutta jo varhain siirryttiin nauhoituksiin. Lienee ilmeistä, että tällaisessa palvelussa automatisointi ja tietokoneen käyttö on luontevaa. Yhtä ilmeistä on, että sovellus on äärimmäisen yksinkertainen tuotettavan puheen osalta, koska vain kellonaikoja tarvitsee tuottaa. Vaikka vuorokaudessa toki on monta eri kellonaikaa, kaikki ne koostuvat *yhdistelemällä* varsin pientä määrää alkeisosasia.

Rautatieasemilla kuulutetaan junien saapumisia ja lähtöjä. Enimmät kuulutukset ovat rutiiniluontoisia ja koskevat aikataulunmukaisia lähtöjä ja saapumisia. Kuulutukset ovat lisäksi vakiomallin mukaisia, esimerkiksi: PIKAJUNA VIISIKYMMENTÄYKSI TAMPEREELLE LÄHTEE RAITEELTA SEITSEMÄN. HYVÄÄ MATKAA! Junia ja niiden lähtöjä on niin paljon, että niiden kuulutusten hoitamiseksi on ilmeisesti kannattanut tehdä automaattinen järjestelmä.

Hieman monimutkaisempi palvelu on käytössä Wienin julkisissa kulkuneuvoissa, joissa jokaisessa bussissa, raitiovaunussa ja metrojunassa on kuulutus kullakin pysäkillä paitsi pysäkin nimestä, myös erilaisista vaihtoyhteyksistä, joihin pysäkiltä on mahdollisuus. Yhdistelmiä

tulee valtava määrä ja reittien tai aikataulujen muuttuessa kuulutuksiin tulee muutoksia moneen paikkaan. Niinpä kuulutuksia varten on ilmeisesti rakennettu automaattinen järjestelmä, jolla tuotettujen äänitysten ansiosta miljoonakaupungin jokaisessa julkisessa liikennevälineessä on ajantasaiset kuulutukset.

Tekstin ääneen lukeminen

Sokeat ja vaikeasti näkövammaiset eivät voi lukea painettua tekstiä sellaisenaan. *Pistekirjoituksella* laaditut erityiset sokeainkirjat ovat kalliita valmistaa ja kopioida. Ne ovat myös kömpelöitä ja tilaa vieviä. Tietokoneen muodossa olevaa tekstiä, (muttei kuvia), sokeat voivat lukea erityisellä mikrotietokoneen lisälaitteella, joka näyttää muutamia kymmeniä merkkejä ruudulla olevaa tekstiä pistekirjoituksena. Tämä keino on kuitenkin aika kallis ja hidas verrattuna siihen, kuinka nopeasti ja vaivattomasti he voisivat vastaanottaa ääneen luettua tekstiä. Näkövammaisten kirjasto tuottaa kyllä kirjoista *äänikirjoja* eli ihmisvoimin ääneen luettuja nauhoituksia, mutta niiden valmistaminen on kallista ja aikaa vievää, joskin tulos on kätevämmän kokoinen kuin pistekirjoituskirja ja helposti kopioitavissa.

Näkövammaisilla on myös käytössään *puhesyntetisaattoreita*, joiden avulla mitä tahansa tekstiä voidaan lukea ääneen koneellisesti. Tiedostona oleva teksti saadaan täten korvin kuultavaksi, mikä on suureksi avuksi. Tässä on hyvä huomata, että sokean korvat vastaavat näkevän silmiä. Näkevälle lukijalle tekstissä ei ole mitään ilmeikkyyttä ja hyvin vähän korostuksiakaan. Sokeakaan ei yleensä halua tyyliteltyä eikä kovin tulkitsevaa lukemista, koska

sellainen tuntuu helposti maneeriselta. Parasta lienee sujuva, *riittävän nopea* selkeästi artikuloitu puhe. *Ohjailtavuus* on myös tarpeen eli ohjelma pitää voida helposti keskeyttää, pyytää uusintaa tarvittaessa hidastettuna tai vaikkapa *kirjaimittain luettuna*. Käytännön sovelluksissa tietokoneohjelmat eivät voi täysin virheettömästi päätellä tekstistä, kuinka erilaiset *lyhenteet, numeroilmaukset tai vierasperäiset nimet* pitäisi tulkita ja lukea. Esimerkiksi lyhenne MM voi tarkoittaa erilaisia asioita (MUUN MUASSA, MILLIMETRI, MAAILMANMESTARUUS jne.). Käyttäjän tulee voida selvittää ongelmakohdat palaamalla tarvittaessa takaisin ja esim. kirjaimittain tapahtuvalla lukemisella.

Sähköpostin lukeminen ääneen puhelimen välityksellä on eräs varhain kehitetty vuorovaikutteinen sovellus, jossa tarvitaan puhesynteesiä. Samoin kuin edellinen, tässäkin on voitava syntetisoida vapaata, rajoittamatonta tekstiä. Sovellus edellyttää myös *vuorovaikutusta* sekä puheentunnistusta, joista enemmän tuonnempana. Itse synteesiosa tarvitsee myös *kielen tunnistusta*, jotta esim. suomenkielisen tekstin joukossa oleva englanninkielinen jakso saataisiin luetuksi ymmärrettävällä tavalla. Tällaisen sovelluksen tulee myös osata hahmottaa pelkän muotoilun perusteella erilaisia jaksoja viestistä ja *viestin pääpiirteittäistä rakennetta*, vaikka viestissä ei olisi nimenomaisia rakennetta ilmaisevia merkintöjä. Monia osia voitaisiinkin tunnistaa sisennysten, rivien alkujen, isojen kirjainten käytön ym. perusteella.

Sähköpostiviestien lisäksi olisi tarkoituksenmukaista tarjota myös muuta normaalisti tekstinä tuotettua materiaalia, esimerkiksi *päivälehtiä* myös puhuttuna versiona sekä näkövammaisille että tavallisille ihmisille, jotka eivät

ole riittävän laajan näytön ja näppäimistön äärellä.

4.2.2 Puhesynteesin menetelmiä

Puhesynteesin toteuttamisessa käytetään erilaisia menetelmiä sovelluksen mukaan. Joissakin on tuotettava hyvin kaavamaista puhetta kuten edellä mainituissa Neiti Aika, junakuulutus ja Wienin julkisen liikenteen kuulutuksissa. Niissä on tarkoituksenmukaista käyttää mahdollisimman pitkiä valmiita osia, jolloin puheen prosodiakin menee paremmin kohdalleen. Juoksevaa tekstiä ääneen luettaessa pitkien valmiiden jaksojen käyttäminen ei ole samassa määrin mahdollista, vaan puhe on tuotettava sananmuodoista ja usein niitäkin pienemmistä osasista.

Leikkaa ja liimaa

Edellä on käsitelty puheen vaihtelevuutta, erityisesti sitä, että äänteet eivät toteudu eri yhteyksissä samalla tavalla ja sitä, että äänteiden keston, painotuksen ja sävelkulun mekanismeja ei tunneta kovinkaan tarkasti. Keinotekoisesti tuotettu puhe kuulostaakin usein konemaiselta ja liian tasaiselta. Todellisen ihmisen nauhoitettu puhe puolestaan on luontevaa — ja *sitä luontevampaa mitä pitempinä yhtäjaksoisina otoksina* sitä toistetaan.

Kuulutuksen osia joudutaan kuitenkin liimailemaan yhteen kaikkien yhdistelmien aikaansaamiseksi ilman, että ihminen joutuu lukemaan yhdistelmiä kokonaisina erikseen. Tällöin on usein vaikeaa saada palasia *sopimaan toisiinsa* luontevasti, jos vaihtoehtoisia osia tallennettaessa esimerkiksi äänenkorkeus vaihtelee. Silloin kokonaisuuden epäluonteva prosodia särähtää kuulijan korvaan.

Toinen yleinen rajoitus leikkaa ja liimaa -menetelmässä on se, että sillä menetelmällä syntetisoitavaa puhetta on äärimmäisen vaikea *muunnella, esimerkiksi nopeuttaa*. Suora nopeutus nostaisi äänen taajuutta, mistä syntyy samanlainen vaikutelma kuin Pikkuoravien kimeässä laulussa. Nopeassa puheessa ei kyse myöskään ole vain tasaisesti lyhyemmistä ja nopeammin toisiaan seuraavista ääntеistä, vaan nopeammassa puheessa *tietyt kohdat redusoituvat enemmän kuin toiset*. Kuulutussovelluksissa ei onneksi muunneltavuus ole olennaista.

Kirjoitetun tekstin laventaminen

Kirjoitettu teksti ei ole tarkoitettu luettavaksi ääneen aivan sellaisenaan, vaan edellyttää yleensä myös jonkinasteista tulkintaa. Vaikka puhesyntetisaattori osaisi ääntää tietyn kielen sanoja, ei se ilman muuta osaa lukea lyhenteitä tai numeromerkintöjä siten kuin ihminen ne lukisi.

Jotkut suomen kielen lyhenteet luetaan *kirjaimittain*, esimerkiksi ATK luetaan AATEEKOO ja jotkut lyhenteet pitäisi avata *sanoiksi, joista ne on lyhennetty*, esim. mittayksiköt, 5 M MATKALLA pitäisi laventaa muotoon VIIDEN METRIN MATKALLA. Monet lyhenteet voi vaihtoehtoisesti joko täydentää pitempään muotoonsa tai sitten lukea lyhennettynä sanana, esim. FIL. KAND. voidaan lukea myös FIL-KAND ja MM-KISAT lukea muodossa ÄM-ÄM-KISAT.

Numeroilmauksiin liittyy muitakin ongelmia, kuin oikean taivutusmuodon päätteleminen tekstiympäristöstä. Esimerkiksi *puhelinnumeroita* ja *päiväyksiä* luetaan eri tavalla kuin hintoja tai muita määriä. Kuulija hämmentyisi, jos kuulisi puhelinnumerona lukuarvon NELJÄSA-

TAAYHDEKSÄNKYMMENTÄKOLME TUHATTA KUUSISATAAKAKSIKYMMENTÄKOLME tai hintana numeroittain luetun YKSI YHDEKSÄN VIISI NOLLA NOLLA. Kieliteknologiaan kuuluva *tiedon eristäminen* pyrkii ratkaisemaan tällaisia tehtäviä (sen lisäksi että tiedon eristäminen tunnistaa henkilöitä, organisaatioita ynnä muita nimettyjä kokonaisuuksia).

Erillinen ongelma on myös *vierasperäisten* (henkilöiden ja paikkojen) *nimien ääntämys*. Valistuneella lukijalla oletetaan olevan pääpiirteittäistä tietoa erikielisten nimien ääntämyksestä. Oletetaan, että SCHUMACHER äännetään suhu-s:llä (š). Taitojen rajat tietysti vaihtelevat: italialainen FISICHELLA saatetaan ääntää aivan oikein FISIKELLA, mutta miten pitäisi ääntää brasilialaisen BARRICHELLON nimi (varsinkin silloin kun hän on italialaisen Ferrarin tallissa). Muutenkin, vierasperäisten nimien ääntämisessä omakielisen tekstin keskellä pitää olla konservatiivinen. Tavoite ei ole mahdollisimman aito alkukielinen ääntämys, vaan *oman kielen* (varovaisesti laajennetulla) äänteistöllä tuotettu likiarvo alkukielisestä ääntämyksestä.

Difonisynteesi

Difoniksi kutsutaan kahden vierekkäisen äänteen puolikkaista muodostuvaa kokonaisuutta, jossa on edellisen äänteen loppupuolisko ja seuraavan äänteen alkupuolisko. Tällaisen difonin käyttökelpoisuus perustuu siihen, että monet äänteet eivät itsenänsä oikeastaan ilmene juuri lainkaan. Esimerkiksi klusiilit K, P, T ovat lähinnä hiljaisuutta, mutta K:n jäljessä oleva A on *erilainen* kuin P:n jäljessä oleva. Itse asiassa tällaiset erot lienevät ainoita, minkä perusteella voimme kuulla (tai kuvitella kuulevamme)

klusiileja eri äänteinä. Difonit kuten KA ovat siten varsin hyödyllisiä yksiköitä, ainakin parempia kuin yksittäisiä foneemeja vastaavat foonit.

Edellisen äänteen loppupuoliskosta ja jälkimmäisen äänteen alkupuoliskosta muodostuvat *difonit* ovat alkeisosasia, joita voi liimata peräkkäin halutulla tavalla. Yksittäisiä difoneja voidaan puolestaan muunnella *kestonsa, intensiteettinsä ja taajuutensa* suhteen, joten tällä menetelmällä pitäisi voida tuottaa luonnollisentuntuista puhetta. Vaihdeltavuuteen liittyy kuitenkin vaatimus määritellä nuo kaikki difonien parametrit joka kohdassa (kun ei voida kopioida suoraan sitä, miten ihminen sen äänsi). Difonien lisäksi voidaan myös käyttää kolmea peräkkäistä äännettä kuvaavia nk. *trifoneja*.

Jos ja kun tiedetään, miten sävelkulku, äänenkestot, tauot ja painotus tulee milloinkin asettaa, saadaan difonisynteesillä tuotetuksi verraten laadukasta puhetta. Jos lähtökohtana on sellainen keskustelujärjestelmä, jossa keskustelua ohjaava komponentti voi ottaa kantaa asiakkaan vuorosanoihin, ei ole mitenkään utooppista liittää difonisynteesin tuotoksiin myös *emootioita eli tunteita*, eli koneen tuotos voidaan saada ilmentymään *hämmästystä, rohkaisua, närkästystä* tai muuta keskustelun onnistumista edistävää tilaa.

Formanttisynteesi

Teoksessa Harrington & Cassidy (1997) on luvussa 7, s. 195–210, kuvattu puhesynteesin historiaa ja formanttisyn-

teesin periaatteet.[4] Formanttisynteesillä tarkoitetaan menetelmiä, joissa puheääni tuotetaan enemmän tai vähemmän täysin synteettisesti moduloiden sitä siihen tapaan kuin, mitä ääntöelimissäkin puhuttaessa tapahtuu.

Formanttisynteesillä tuotettavaa puheääntä on helpompi *muunnella* kuin eriasteisista osasista koottavaa puheääntä. Erityisesti silloin, kun puhetta ei tuotetakaan tekstistä, vaan tietokoneohjelman tuottamasta loogisesta tiedosta, on muunneltavuudella arvoa. Tunteiden sisällyttämisellä arvoa, jos esimerkiksi *vuorovaikutteinen* keskusteleva ohjelma haluaa esittää ihmetystä tai epäuskoa käyttäjän epätavallisen vastauksen vuoksi tai kenties rohkaista liiankin hitaasti ja varovaisesti etenevää käyttäjää.

4.2.3 Puhesynteesin ongelmia

Puheen prosodiaa *ei tunneta* kovin tarkasti, eikä siis aivan tiedetä minkälaisilla parametreilla puhe pitäisi syntetisoida, jotta se kulloinkin kuulostaisi luontevalta.

Jos luetaan valmiiksi kirjoitettua tekstiä ääneen, pitäisi teksti osata *jäsentää* ja tulkita kyllin hyvin, jotta prosodia saataisiin paikalleen.

Paremmassa tilanteessa ollaan silloin, kun tietokoneella ollaan *generoimassa* vastausta esimerkiksi tietokantahaun tuloksista, jolloin yleensä ei ole rakenteellisia monitulkintaisuuksia eikä ongelmia siinä, mitä lyhenteet tai luvut tarkoittavat. Lisäksi virkkeen sisäiset lausekkeiden rajat ovat tällaisissa sovelluksissa ilman muuta selvillä.

Yksi puhesynteesi ei ehkä tyydytä kaikkia. Sokeille tai

[4] Wikipedian *Speech synthesis* -artikkelissa on myös selostettu formanttisynteesiä

näkövammaisille puhesynteesi voi olla vain silmän korvike ja heille sopii nopea ja pelkistetty synteesi. Toisaalta tavallinen eli näkevä käyttäjä närkästyy ehkä helpommin, jos luettu teksti ei ole totutun normin mukaista.

4.3 Puheentunnistus

Puheentunnistukselle on suuri tarve, kuten johdannossa jo havaittiin. Monissa tilanteissa laitteita tai järjestelmiä olisi luonnollista komentaa puheen avulla. Tilausta olisi enemmällekin, kuin mihin tekniikka pystyy.

4.3.1 Puheentunnistuksen sovellukset

Puheentunnistuksen käytännön sovellukset ovat kaikki melko uusia ja jotkut vielä vakiintumattomia ja enimmäkseen kaukana tieteiselokuvien puhetta sujuvasti ymmärtävistä järjestelmistä. Kuitenkin puheentunnistus on tullut osaksi joitakin jokapäiväisiä sovelluksia.

Puheella komennettavat laitteet

Markkinoilla on *matkapuhelimia*, joita voi ohjata puheella. Kännykälle voi opettaa kymmeniä nimiä siten, että painamalla sopivaa nappia ja sanomalla jonkin näistä nimistä laite soittaa toivottuun numeroon.

Paljon on puhuttu siitä, että *autossa olevia laitteita* pitäisi voida tarvittaessa komentaa *puheohjauksella*. Yleisesti ottaen tarvetta puheohjaukselle lienee muuallakin tilanteissa, missä käyttäjän *kädet ovat kiinni* tai katseen pitäminen ohjattavan laitteen nappuloissa tai säätimissä on hankalaa tai vaarallista. Nyttemmin on toteutettu esim.

puheohjattavia hammaslääkärin tuoleja, kaiketi juuri siksi, että hammaslääkärin kädet tarvitaan instrumentteihin, eivätkä olisi samalla käytettävissä laitteiden säätöön tms.

Yksinkertaiset puheohjatut palvelut

Lukemattomissa firmoissa ja toimistoissa on *puhelinkeskuksessa* joku, jolta soittajat pyytävät nimellä, kenelle haluavat yhdistettäväksi ja keskus sitten *yhdistää puhelun* toivomuksen mukaan. Tähän ei kaiketi tarvita ihmistä, jos keskuksessa on laitteisto, joka tunnistaa puheesta halutun nimen ja suorittaa automaattisesti yhdistämisen. Tällaiset sovellukset ovat melko helposti toteutettavissa nykytekniikalla sen vuoksi, että luettelossa olevia nimiä on yleensä melko rajallinen määrä ja etunimen ja sukunimen yhdistelmät eroavat hyvinkin selvästi toisistaan.

Numerotiedotuksessa asiakas kyselee kaupungin tai laajemman alueen piirissä asuvan ihmisen nimen perusteella hänen puhelinnumeroansa. Perinteisesti tätä on hoidettu siten, että palvelussa on ihmisiä, joilla on luettelot (joko perinteisinä kirjoina tai tietokantana) saatavilla ja asiakkaan toivoma henkilö etsitään niistä luetteloista. Tehtävän automatisointi on vaativampi kuin edellinen, sillä tarjolla on mahdollisia nimiä huomattavan suuri määrä. Tällaisia sovelluksia alkaa kuitenkin olla markkinoilla.

Sanelusovellukset

Tietokone automaattisena sihteerin tai *konekirjoittajan korvikkeena* on ollut kauan mukana tulevaisuuden toimiston visioissa. Sanelukoneen sijasta puhe muuttuisi sellaisessa suoraan tekstiksi automaattisen puheentunnistuksen

menetelmillä.

Visio on muuttunut tietyiltä osiltaan todellisuudeksi ja käytännössäkin mahdolliseksi. Eri valmistajat kuten Philips ja IBM ovat tuottaneet kaupallisia ohjelmia, joilla voidaan muuttaa *puhetta tekstiksi*. Nyttemmin Nuanceniminen firma ja Google ovat astuneet näiden tilalle.[5] Useimmat näistä ohjelmista on harjoitettava kullekin puhujalle etukäteen, mutta sen jälkeen puhuja voi käyttää laajaa sanastoa, jonka ohjelma tunnistaa. Tehtävä on vaikea ja tällä hetkellä saatavilla olevien ohjelmien laatu vaihtelee sekä kielen että tuotteen mukaan. Kokeile esim. Googlella etsimällä termeillä SPEECH RECOGNITION DICTATION löytääksesi verkosta tietoa näistä ja muista tuotteista.

Vuorovaikutukseen perustuvat puheohjatut palvelut

Kysyttäessä *aikatauluja* tai bussiyhteyksiä tarvitaan usein kohtuullista sanastoa ja nimenomaisesti myös vuorovaikutteisuutta. Sanastona saattaa olla tietyn kaupungin katujen, rakennusten ja muunlaisten paikkojen nimet sekä pienempi joukko keskusteluun tarvittavat muita sanoja. Ihmiset saattavat höystää kysymyksiään monenlaisilla lisukkeilla, mutta paikkoihin liittyvät sanat tai muut tärpit yleensä kertovat, mistä kulloinkin on kyse. *Lentolipun tai elokuvalipun* varaaminen on vaikeudeltaan samatapainen sovellus ja siten ilmeisesti täysin toteutettavissa.

[5] ks. esimerkiksi:
http://www.nuance.com/for-individuals/by-solution/speech-recognition/

4.3.2 Puheentunnistuksen menetelmiä

Puheentunnistusta selvitellään teoksessa Mitkov (2003) luku 16, (Speech Recognition), s. 305–322. Yksittäisiä komentosanoja voidaan tunnistaa ihan vain vertaamalla joitakin puheesta laskettuja parametreja mallisanoihin ilman, että yritetään eristää yksittäisiä äänteitä tai muutakaan kielellisesti motivoitua käsittelyä.

Varsinainen puheentunnistus perustuu useampaan vaiheeseen, aluksi esim.:

- Analoginen puhesignaali *digitoidaan* eli ääniaaltojen värähtely muutetaan sarjaksi lukuarvoja, jotka kuvaavat hyvinkin tarkasti ääntä.

- Digitoitua signaalia käsitellään esimerkiksi erittelemällä eri taajuuksilla oleva energia käyttäen FFT (Fast Fourier Transformation) -algoritmia *spektrin* laskemiseksi. Spektri on kunakin muutaman kymmenen millisekunnin aikaviipaleena riippuvainen äänteen laadusta. Spektristä tarvitaan tunnuslukuja, jollaisina voidaan käyttää ns. cepstrejä, jotka kuvaavat signaalin aikaviipaleen spektrin energiajakautuman huippuja tiivistetyssä muodossa.

- Näistä tunnusluvuista muodostetaan kyseistä aikaviipaletta kuvaava vektori.

Puheentunnistuksessa on ongelmana se, että mikään *yksittäinen* kohta puheessa ei ole varmuudella tunnistettavissa. Jotkut kohdat erityisen epävarmoja ja toiset varmempia. Siinä missä yksittäiset viipaleet ovat epävarmoja, ovat *viipaleiden sarjat* astetta varmempia. Muutamien

peräkkäisten viipaleiden pitää yleensä kuulua samaan äänteeseen (tai difoniin) eikä poukkoilu hyvin nopeasti äänteistä toiseen ole mahdollista, vaan äänteen vaihtumisia voi tapahtua vain tietyissä rajoissa. Peräkkäisten tulkintojen pitää olla keskenään johdonmukaisia ja niiden pitää kokonaisuutena muodostaa jokin mahdollinen kokonaisuus (esim. sana). Sen vuoksi prosessi jatkuu käyttäen valistunutta arvausmenettelyä.

- Eri difoneille ja niiden äänestä lasketuille vastaaville vektoreille on muodostettu ns. *Markovin piilomalli* (HMM eli Hidden Markov Model), joka on laadittu äärellisen tilasiirtymäverkon muotoon. Tunnistettavaa ääntä kuvaavien vektorien sarjaa verrataan näihin verkkoihin etsien sellaisia reittejä, jotka olisivat kokonaisuutena todennäköisiä. Reitit muodostavat äänteitä ja edelleen sanoja tai fraaseja, jotka on etukäteen ns. kielimallilla määritelty.

Vaihtoehtoisia menetelmiä on eri vaiheille olemassa. *Kielimalli* voi olla yksinkertainen luettelo, tilasiirtymäverkko tai jotakin muuta. Signaalin viipaleita kuvaavien vektorien luokitteluun on käytetty HMM:n sijasta myös hermoverkkoja ja muita menetelmiä.

4.3.3 Puheentunnistuksen ongelmia ja mahdollisuuksia

Puheentunnistuksen nykytaso mahdollistaa:

- joko *laajalle puhujajoukolle rajattuun sanastoon perustuvia* tai

- *yhdelle puhujalle sovitettuja laajaan tai rajoittamattomaan sanastoon* perustuvia sovelluksia.

Näistä edellinen on toteutettavissa virheettömämmin kuin jälkimmäinen. Rajoittamattoman tekstin syöttämisessä yhdeltäkin ihmiseltä on varauduttava *tarkisteluun ja korjailuun*, jos mielitään suhteellisen virheetöntä tulosta.

Vaikka monissa tilanteissa puheen käyttäminen on mahdollista ja luontevaa, *puhekäyttöliittymä ei sovi kaikkialle*. Esimerkiksi toimistoympäristössä näyttöruudulta lukeminen ja tekstin näppäily tai hiiren käyttäminen puolustanevat asemaansa ainakin sen vuoksi, että ääneen puhuminen häiritsee toisia tai kaikkia asioita ei haluta toisten kuuluville. Perinteisesti itsekseen puhumista on pidetty arveluttavana, mutta ehkä kännyköiden aikana tämä vieroksunta lievenee osittain. Riittävän kokoinen näyttö ja hiiri sallivat monin tarkoituksiin paremman ja nopeamman ratkaisun kuin puhe.

Puheentunnistustehtävää vaikeuttaa puheen vaihtelevuus saman puhujan puheessa eri kerroilla ja aivan erityisesti erot eri puhujien puheen välillä. Toistaiseksi ei oikein tunneta, mitkä fysikaaliset ominaisuudet puheessa tekevät siitä *ihmisen korvalle* tunnistettavaa. Tietysti kelpaisi sekin, jos löydettäisiin puheesta muita piirteitä, joiden avulla puheen oikeat sanat voitaisiin erottaa yhtä virheettömästi kuin ihmiskorva tai kenties virheettömämminkin.

On spekuloitu, että tulevaisuudessa voisi olla sellainen *älykäs koti* jossa katkaisijat voivat toimia puhekomennoilla ja jossa voisi kysellä mahtoiko liesi jäädä päälle tai onko keittiön ikkuna auki. Tällaisella voisi olla todellista arvoa ainakin vanhuksille ja vammaisille.

Puhekäyttöliittymän yleistyminen edellyttää varmaan *uusien tapojen* muotoutumista puheäänen käyttämiseen. Matkapuhelinten vakiintumisen myötä tämä muutos ja sopeutuminen varmaan on jo käynnissä.

Luku 5

Luonnolliskielinen vuorovaikutus tietokoneiden kanssa

Aiempana on useassa yhteydessä viitattu siihen, että eri sovelluksiin liittyy tarvetta vuorovaikutukseen ihmisen ja koneen välillä. Ihmistenkin välisestä kielen käytöstä suuri osa on vuorovaikutteista. Kuulija ymmärtää vain osittain ja ilmaisee suoraan tai epäsuorasti, miten keskustelua pitäisi jatkaa. Epävarmuus ja moniselitteisyys vain korostuvat, kun tietokone on keskustelun toisena osapuolena.

5.1 Taustaa

Keskustelu tietokoneen kanssa, siis ihmisen ja tietokoneen välinen luonnollisella kielellä tapahtuva vuorovaikutus, on monella tavalla kiehtova haastava asia. Sitä voidaan kuitenkin tarkastella eri tahoilta, eikä kyse ole vain yhdestä

asiasta.

5.1.1 Miksi ylipäätänsä keskustella koneen kanssa?

Tietokoneen käyttämisestä on tullut miljoonien arkipäivää, mutta koneen kanssa ei varsinaisesti keskustella, vaan sitä *komennetaan* näppäimillä, hiirellä osoittamalla tai kosketusnäytöllä. Useita arkipäivän sovelluksia voidaankin ohjata aivan hyvin näinkin mekaanisesti. *Näppäimistö* on luonteva tapa syöttää tekstiä. Tekstin muotoilemiseksi halutunlaiseksi dokumentiksi on totuttu käyttämään *hiirtä* sekä arvioimaan lopputulosta ruudulla sopivien asetteluiden ja kirjasintyylien valitsemiseksi. Yhtä lailla sähköpostin lukeminen on ohjattavissa näppäimillä ja hiirtä napsauttamalla, eikä verkkoselaimenkaan käyttö ole näillä keinoin mitenkään ongelmallista.

Luonnollisen kielen käyttäminen vuorovaikutuksessa ei siten ole ilmeistä, vaan tarve siihen voi syntyä muutamista *syistä*: (1) käyttäjä on poissa työpöytänsä äärestä ja hänellä on mukanaan vain niin *pieni laite*, että siinä ei ole kunnon näppäimistöä, hiirtä eikä isoa ruutua, (2) käyttäjällä on työtehtäviensä takia *kädet sidottuna muuhun* työhön, minkä vuoksi niitä ei voi käyttää tai (3) sovellus on aidosti *mutkikkaampi* ja erilainen kuin yllä mainitut vakiosovellukset.

Ensimmäinen vaihtoehto koskee luonnollisesti *matkapuhelinta*, joka on pieni ja sille on helpointa puhua ja kuunnella sieltä puhetta. Vaikka matkapuhelimissa on näyttö, se on pakostakin pieni verrattuna työpöydällä käytettäviin laitteisiin ja näytön osoittaminen on myös han-

kalampaa tai ainakin epätarkkaa. Jos sovellusta ryhdytään ohjaamaan puheen avulla, järjestelmän täytyy olla vuorovaikutteisempi kuin näppäimistön ja hiiren kanssa, sillä puheentunnistus on jonkin verran satunnaisesti epävarmaa. Sen vuoksi on ainakin silloin tällöin *tarvetta varmistaa*, onko ihmisen puhuma repliikki tulkittu oikein.

Vuorovaikutteisuuden voi kääntää myös *vahvuudeksi*. Ihmiset eivät nimittäin aina ole eivätkä haluakaan olla niin systemaattisia ja tarkkoja kysymyksissään ja vastauksissaan, kuin perinteinen tietojenkäsittely edellyttäisi. Ihmiset turhautuisivat, jos heidät pakotettaisiin etenemään puheessaan yhtä kaavamaisesti kuin, mitä heiltä vaaditaan graafista käyttöliittymää käytettäessä.

Eräs tarve vuorovaikutteisuuteen johtuu siis yksinkertaisesti puheen käyttämisestä koneen kanssa eli puheentunnistuksen epätarkkuudesta. Toinen mahdollinen peruste luonnollisen kielen käyttämiseen vuorovaikutuksessa koneiden kanssa on tehtävän laatu. Jos mahdollisuuksia on erittäin paljon, ei graafisella liitännällä ehkä ole helppo tuoda niitä tarjolle. Monet tehtävät ovat käyttöliittymältään hyvin yksinkertaisia kuten Internetin hakukoneet, jotka toimivat usein melkeinpä liian hyvin. Esimerkiksi kun käytetään harvinaisempia henkilöiden nimiä tai paikannimiä, voidaan haluttu sivu löytää suoraan tällaisen nimen avulla. Kaikissa tapauksissa löytäminen ei voi olla näin helppoa, vaan etsimisen täytyisi kulkea *vaiheittain ja vähitellen tarkentaen ja rajaten* hakua.

Oman erityisen kenttänsä muodostavat *vammaissovellukset*. Sokealle tai näkövammaiselle ei ruudusta välttämättä ole apua, mutta puhe toimii ongelmitta kumpaankin suuntaan.

5.1.2 Keskustelu puhuen vai kirjoittaen?

Ihminen voi keskustella tietokoneen kanssa joko tekstin tai puheen muodossa. Teknisesti olisi helpompaa toteuttaa vuorovaikutusta, jossa ihminen kirjoittaa vuorosanansa näppäimistöllä ja lukee koneen vastaukset tekstinä ruudulta. Käyttäjälle näppäily voi puolestaan olla vaivalloisempaa, vaikka toisaalta viesti menee koneelle tarkasti juuri sellaisena, kuin se on näppäilty.

Puheentunnistus ja puhesynteesi ovat toki teknisesti vaivalloisia toteutettavaksi. Vaikka puheen käyttö kommunikointiin olisikin tehty mahdolliseksi, ihmiset eivät silti *aina* edes haluaisi puhua koneelle. Esimerkiksi toimistossa ja kotonakin on muita kuulemassa, jolloin puhekäyttöliittymän käyttö olisi kiusallista. Moni tekisi mieluummin työtään omassa rauhassaan ilman, että muut kuulevat mitä ollaan tekemässä.

Keskustelu puhumalla on kuitenkin luontevampaa kuin kirjoittamalla (vaikka ns. chattaily ehkä lähestyy puhuttua keskustelua). Puheella on erityisen helppo antaa *täydentäviä, varsinkin lyhyitä ohjeita, lisäkysymyksiä, myöntäviä ja kieltäviä vastauksia* ym.

Puhuminen on aina vaihtelevaa, puhesignaali vähän *epämääräistä* ja siitäkin syystä puheentunnistus on jonkin verran epätarkkaa. Puhetta ei myöskään *suunnitella* niin huolellisesti kuin kirjallista esitystä, joten usein sanotaan vain osa tarvittavasta asiasta ja sekin ehkä monitulkintaisesti. Puhekeskusteluun liittyy siten olennaisesti tarvetta antaa erilaista *palautetta ja herätettä* tarvittavan keskustelun kuljettamiseksi tarkoituksenmukaisesti eteenpäin.

5.1.3 Koneen ja ihmisen yhteispeli

Ollakseen ihmiselle miellyttävää, keskustelun tulee edetä johdonmukaisesti ja sujuvasti. Erityisen tärkeää on sopiva *annostelu*. Koneen tulee antaa ohjeita, tietoja ja vastauksia *kohtuullisen kokoisia määriä kerrallaan* ja tarpeen mukaan.

Vaikka puheentunnistus on nykyisellään aina jonkin verran epätarkkaa, tätä epävarmuutta ei voida ratkaista pelkästään sillä, että käyttäjän *jokaisen vuorosanan* jälkeen kone tarkistaa, onko se kuullut oikein. Käyttäjät nimittäin hermostuisivat tällaisesta ja pian luopuisivat koko yrityksestä käyttää palvelua, joka ei tunnu pääsevän eteenpäin. Varmistuksen tuleekin usein olla *epäsuoraa* siten, että kone ei suinkaan kysy, sanoiko käyttäjä tietyllä tavalla, vaan varmistuksen voi myös sisällyttää epäsuorasti vuorosanoihin.

Vuorovaikutusta on tapana kuvata siten, että meillä on kaksi keskustelevaa osapuolta ja kummallakin oma *sisäinen tilansa*, joka vähitellen muuttuu keskustelun kuluessa. Osapuolet eivät tiedä toistensa tiloja tarkalleen, mutta muodostavat oman likimääräisen *arvionsa* toisen osapuolen tilasta, johon ne perustavat ratkaisunsa siitä, mitä seuraavaksi kannattaisi sanoa. Sisäiseen tilaan voi kuulua kaikenlaista: keskustelun historiaa, tähän mennessä koottua tietoa, käsitystä keskustelukumppanin tilasta sekä erilaisia uskomuksia ja tietoja ympäröivästä maailmasta tai keskustelun kohteena olevasta teemasta.

5.2 Vuorovaikutussovelluksia

Seuraavassa tarkastellaan sitä, millaista luonnollista kieltä käyttävä vuorovaikutus koneen kanssa voisi olla ja millaisia tarkoituksia varten sitä on kehitelty. Vuorovaikutteisissa sovelluksissa, samoin kuin muissakin, vain osasta mahdollisia sovelluskohteita on jo saatavilla valmiita ja kattavia sovelluksia. Useimmat kohteet odottavat toteuttajaansa.

5.2.1 Jutustelu koneen kanssa

Tietokoneohjelmoinnin teoreettisen kehittäjän Alan Turingin mielessä liikkui jo ennen varsinaisten kunnollisten tietokoneiden rakentamista kysymys siitä, voisiko tietokone keskustella kuten ihminen. Häneltä on peräisin (v. 1950 julkaistu) nyttemmin *Turingin testin* nimellä kulkeva koe, jolla pyritään arvioimaan keskustelevien tietokoneohjelmien hyvyyttä. Testi koskee nimenomaan vapaampaa, (mutta tiettyyn aihepiiriin ehkä rajautuvaa) keskustelua. Tietokoneen keskusteluohjelman hyvyyttä arvioidaan sillä, kuinka varmasti tai todennäköisesti ulkopuolinen kyselijä voi *erottaa oikean ihmisen tästä ohjelmoidusta koneesta*. Mittarina voidaan käyttää aikaa, jonka tällainen kyselijä tarvitsee saavuttaakseen tietyn onnistumisen asteen arvauksessa.

Turingin testin perusteella on vuodesta 1991 järjestetty vakavahenkisiä kilpailuja, joissa jaetaan joka vuosi ns. *Loebnerin palkinto* Turingin testissä parhaiten selviytyneelle tietokoneohjelmalle.[1]

Kauan ennen näitä vuosittain järjestettäviä kilpailui-

[1] Ks. http:

ta tehtiin muutamia samanhenkisiä ohjelmia, joiden sanotaan harhauttaneen käyttäjiä luulemaan ohjelmaa inhimilliseksi keskustelukumppaniksi. Tunnetuin näistä on Joseph Weizenbaumin *ELIZA* vuodelta 1966, josta on nykyäänkin on verkossa saatavissa uudelleen toteutettuja demoja ja lukuisia muunnelmia. ELIZA jäljitteli Rogerin oppisuunnan *psykoterapeuttia*. Käyttäjän odotettiin kertovan sille omista huolistaan, mutta ohjelman tuli vain hienovaraisesti auttaa ihmistä etenemään keskustelussa. ELIZA oli varsin yksinkertainen tietokoneohjelma ja sen menestys perustui enemmän siihen, että yhteistyöhön halukas ihminen tuotti omissa repliikeissään pääosan keskustelun materiaalista, jota ohjelma sitten syötti takaisin vähän valikoiden ja muunnellen. Vahva illuusio ymmärtävästä ohjelmasta syntyi vain ihmiskäyttäjän mielessä. ELIZAn mallin mukaan on tuotettu paljon samantapaisia ohjelmia,[2].

Toinen klassinen dialogiohjelma oli psykiatri Kenneth Colbyn vuonna 1972 laatima PARRY-niminen ohjelma. PARRY liittyi myös mielenterveyteen siten, että se näytteli paranoiaa sairastavaa keskustelun osapuolta.[3]

5.2.2 Laitteiden ohjaaminen ja sanelu

Autoa ajettaessa pitäisi *silmien* olla tarkkaavaisesti seuraamassa liikennettä ja *käsien* kiinni ohjauspyörässä, mutta joitakin autossa olevia laitteita, kuten soittimia, lämmittimiä jne. pitäisi silti voida käyttää. Hammaslääkärin tuo-

---

//www.loebner.net/Prizef/loebner-prize.html

[2]ks. esim. Wikipedia, http://en.wikipedia.org/wiki/ELIZA ja edellisen alaviitteen verkko-osoite.

[3]ks. esim. http://en.wikipedia.org/wiki/PARRY

lissa on myös paljon toimintoja, joita ei voi käsin käyttää samalla, kun molemmat *kädet ovat kiinni* instrumenteissa. Tämänkaltaisiin tarkoituksiin puhekäyttöliittymä on mahdollinen ja tarkoituksenmukainen. Liitännässä tarvittava kieli voi olla hyvin redusoitua ja koostua rajallisesta määrästä etukäteen tiedettyjä komentoja. Laitteen ei tarvitse yksikertaisimmillaan oikeastaan muuta kuin toimia komennon mukaan. Vaativammissa sovelluksissa toki tarvitaan palautettakin.

Tekstinkäsittelyäkin voidaan periaatteessa tehdä pelkästään äänen avulla. Silloin käyttäjän täytyy voida suorittaa kaikki muotoilua ja korjailua koskevat komennot *puheen avulla*. Näiden toimintojen ohjaamista on toteutettu siten, että muuten tavanomaista tekstinkäsittelyohjelmaa komennetaan sanallisilla komennoilla, jolloin palaute tulee ruudulla näkyvinä valikkoina samaan tapaan kuin näppäimiäkin käytettäessä. Siten yksittäiset hiiren napsautukset tai näppäilyt on vain korvattu saman toiminnon laukaisevalla komentosanalla. (Tekstinkäsittelyohjelman käyttäminen tällä tavoin ei toki ole nopeampaa eikä sujuvampaa kuin näppäimistöllä ja hiirellä, mutta erikoistapauksissa ehkä tarkoituksenmukaista.)

Laitteiden ja ohjelmien ohjaaminen voi tietysti olla tätä paljon monimutkaisempaa ja hienojakoisempaa, jolloin se muistuttaa tuonnempana lueteltavia sovelluksia.

Tarkastellaan vielä esimerkiksi sitä, miten *sokea voisi lukea päivän lehden*. Puhesynteesillä tekstin voi saada ääneksi, mutta harva haluaisi tai ehtisi lukea koko lehden alusta loppuun. Sivujen silmäilyn vastineeksi tarvitaan paljonkin keskustelua. Koneen pitäisi antaa mahdollisuus edetä halutussa järjestyksessä ja valikoiden. Näin voi tehdä

nykyisissäkin digilehdissä, mutta näkövammaiselle ei voi lukea ääneen kaikkia ruudulla näkyviä vaihtoehtoja, vaan käyttäjän *omin sanoin* esittämiin pyyntöihin pitäisi reagoida. Ehkä isommista uutisista voisi tarjota otsikot ja antaa käyttäjän pyytää lisää niistä, jotka kiinnostavat. Jotkut kohdat, esimerkiksi *taulukot* ovat tekstissä oma haasteensa vuorovaikutukselle. Ei ole aivan selvää, miten puheen avulla voidaan välittää kuulijalle käsitys siitä, mitä taulukossa on nähtävissä. Tuskin yksi taitavakaan läpilukeminen voi riittää taulukon hahmottamiseen, vaan sellaisia varten täytyisi kai kehittää oma dialoginsa, jossa käyttäjä voi kysellä useilla tavoilla taulukon sisältöä.

5.2.3 Tietokantaliitäntä ja sen kaltaiset sovellukset

Eräs vanhimmista vuorovaikutteisista sovelluksista on tietokantakysely. Siinäkin ihminen hoitaa kielen käytön ja kone vastaa tuloksilla. Ensimmäinen tämän lajin ohjelma lienee William Woodsin ATN-kielioppiin (Augmented Transition Networks) perustuva LUNAR. Siinä oli taustalla Apollo-ohjelman *kuukivien analyysien tietokanta*, johon saattoi kohdistaa englanninkielisiä kysymyksiä. Vastaukset olivat tyypillisesti lukumääriä tai näytteiden numeroiden luetteloita.[4]

Samaan tyyppiin tietokantojen käyttöliittymien kanssa voisi lukea sellaisiakin ohjelmia, joissa ei ole tietokantaa, vaan ohjelmallista tietämystä, esim. aritmetiikasta. Esimerkiksi Daniel Bobrowin v. 1968 raportoima STUDENT

[4] ks. esim. http://en.wikipedia.org/wiki/Question_answering_system

-niminen ohjelma osasi analysoida englanninkielisiä virkkeitä, joissa oli *aritmeettisia suhteita*. Ohjelma osasi tehdä niistä yhtälöitä, jotka se ratkaisi. Kokeilkaapa itse ratkaista esimerkiksi seuraavaa, jonka STUDENT tiettävästi osasi:

> Mary is twice as old as Ann was when Mary was as old as Ann is now. If Mary is 24 years old, how old is Ann?

STUDENT-ohjelma luultavasti ohitti syötteestä kaiken, mitä se ei voinut tunnistaa, sillä se oli varautunut tunnistamaan vain sellaisia osia ja jaksoja, joista voitiin muodostaa yhtälö, esim. M2 = 2 * A1 (Maryn ikä nyt on 2 * Annin ikä silloin), M2 = 24 (Mary on nyt 24-vuotias), M1 = A2 (Mary oli silloin yhtä vanha kuin Ann nyt), M2 - M1 = A2 - A1 (yleistä tietoa siitä, että he vanhenevat samaa tahtia). Neljä yhtälöä, joissa yhteensä neljä tuntematonta muodostaa tietokoneohjelmalla helposti ratkaistavan yhtälöryhmän.

5.2.4 Dialogiin perustuvat palvelut

Tyypillisiä puheohjattuja sovelluksia ovat erilaiset *aikatauluneuvonnat* ja järjestelmät, joiden avulla voi *tilata lippuja*. Näistä eräs vanhimmista, GUS-niminen ohjelma, jäljitteli lentolippua myyvää matkatoimiston virkailijaa. Asiakkaan oletettiin haluavan ostaa lentolipun ja ohjelman tehtävänä oli kysellä asiakkaalta lippua varten tarvittavat tiedot kuten mistä lähdetään, minne lennetään, lähtöpäivä ja haluttu lähdön kellonaika jne. Tyypillisesti asiakas saattoi oma-aloitteisesti vastata enempäänkin kuin, mitä kysyttiin tai vastata kysymykseen, jota ei ollut vielä edes

esitetty. Ohjelma keräsi tietoja ja kun kaikki tarvittavat tiedot oli saatu kokoon, lippu oli valmis myytäväksi, vrt. Bobrow et al (1977).[5]

5.3 Keskustelujärjestelmien menetelmät

Vuorovaikutteisten luonnollisen kielen keskustelujärjestelmien toteutukset perustuvat erilaisiin malleihin riippuen siitä, kuinka vapaaksi tai kiinteäksi keskustelun kulku rajataan. Toisiin sovelluksiin riittää yksinkertaisempi malli, mutta toisissa taas on varauduttava haarukoimaan kyselemällä, mitä käyttäjä oikeastaan haluaa.

5.3.1 Äärellistilainen vuorovaikutus

Yksinkertaisimmillaan ihmisen ja koneen vuorovaikutus puheen avulla voi olla melko pelkistettyä. Esimerkiksi puheohjatulle puhelinvaihteelle sanotaan vain halutun henkilön nimi, jonka jälkeen vuorovaikutus on yleensä ohi ja vaihde yhdistää puhelun tuolle henkilölle.

Äärellistilainen keskustelujärjestelmä pystyy tähän ja hieman yleisempään käsittelyyn, jossa keskustelu etenee *yhden askeleen kerrallaan*. Toimintaa voidaan havainnollistaa *äärellistilaisella automaatilla*, jonka siirtymistä tilasta toiseen kyseiset puheviestit ohjaavat. Yhdellä kertaa käyttäjän odotetaan sanovan yhden asian, ei enempää eikä mitään muuta. Jos siis kysytään, minne halutaan matkustaa, käyttäjän ei pidä tässä mallissa kiiruhtaa asioiden

[5] Artikkeli löytyy verkostakin, lokak. 2012 osoitteesta `http://nlp.stanford.edu/acvogel/gus.pdf`

edelle ja kertoa lisäksi, milloin haluaisi olla perillä. Riippuu sovelluksesta, kuinka hyvin *käyttäjät* voivat sopeutua tällaiseen *kurinalaiseen* vuoropuheluun. Puhelinvaihteen kanssa ei varmaan tule isoja ongelmia, vaikka sovelletaan näin yksinkertaista dialogin mallia. Sen sijaan esimerkiksi matkalippujen tilaamisessa käyttäjän lienee vaikeaa sopeutua näin suoraviivaiseen kaavaan, koska matkan koostaminen on aika mutkikas tehtävä. Kun käyttäjä itsekin vasta sommittelee ja suunnittelee matkaansa, ei hänen ehkä ole helppoa vastata suoraan kysymykseen. Hän ehkä ei itsekään vielä tiedä milloin haluaisi lähteä matkaan, mutta tietää kyllä milloin pitäisi olla perillä.

Äärellistilaiseen malliin liittyvä puheentunnistus on hyvin suoraviivaista: tilasiirtymät liittyvät kiinteisiin sanoihin tai fraaseihin, joiden vaihtoehdot ovat etukäteen tiedossa. Käyttäjän vuorosanaa verrataan näihin etukäteen varastoituihin vaihtoehtoihin ja valitaan niistä se, jota käyttäjän repliikki todennäköisimmin vastaa. (McTear, 2002, s. 92–93, kohta 2.1 ja s. 96–98, kohta 3.1)

5.3.2 Kehyksiin pohjautuva vuorovaikutus

Kehyksiin (frame) *perustuva vuorovaikutus* pyrkii hallitsemaan kattavammin koko keskustelutehtävän etenemisen. Tyypillisesti *tarvitaan joukko tietoja*, esim. lentolipun ostamista varten minne matkustetaan ja milloin sekä palataanko takaisin ja milloin. Käyttäjän puheesta pyritään noukkimaan esille jaksoja, jotka täyttäisivät tätä kehystä eli lokerikkoa. Etsitään siten esim. paikannimiä sopivan preposition kanssa tai sopivassa sijamuodossa taikka vastaavasti päivämääriä ja kellonaikoja sekä lähtemistä tai

saapumista koskevia ilmauksia (concept spotting). Niinpä suomenkielinen käyttäjän vuorosana MINUN OLISI TARKOITUS OIKEASTAAN MATKUSTAA TUKHOLMAAN KOKOUKSEEN, JOSSA MINULLA ON ESITELMÄ voi järjestelmän kannalta redusoitua muotoon … TUKHOLMAAN …. Parhaassa tapauksessa yhdestä vuorosanasta voidaan saada irti useampikin hyödyllinen osanen.

Kehyksen asteittainen täyttyminen ja vielä *vajaana olevat lokerot ohjaavat järjestelmän palautetta*. Järjestelmän tavoitteena on saada käyttäjä antamaan vielä puuttuvat tiedot. Koneen vuorosanat voivat silloin olla joustavan tuntuisia käyttäjän on helppo sopeutua dialogin kulkuun, ks. (McTear, 2004, 93–94, kohta 2.2 ja s. 98–100, kohta 3.2).

5.3.3 Agentteihin perustuva vuorovaikutus

Edellä mainitut keskustelujärjestelmien tyypit ovat molemmat tarkkarajaisia sen suhteen, minkälaisiin repliikkeihin ne reagoivat. Kumpikin laji tietää vain siitä tehtävästä, jota varten ne nimenomaan on laadittu. Agenttipohjaisten järjestelmien ajatellaan toisaalta olevan itsenäisempiä ja toisaalta tuntevan oman tehtävänsä lisäksi jossakin määrin ympäristöään. Agentit voivat myös olla tietoisia toisista agenteista, jotka tekevät toisenlaisia tehtäviä. Kukin agentti pyrkii vastaamaan suoraan tai tekoälyllä päättelemällä omaa ympäristöään koskeviin kysymyksiin. Kun agentille tulee kysymys, joka ei kuulu sen omaan ympäristöön, agentin oletetaan tekevän valistuneita arvauksia siitä, mikä toinen agentti voisi olla asiakkaalle kenties hyödyllinen. Karkeasti ottaen kysymys on siitä, että tällaisen järjestel-

män agentti pyrkii olemaan hyödyllinen silloinkin, kun agentilta kysytään asioita, jotka eivät sille kuulu tai johon sitä ei ole suoraan ohjelmoitu vastaamaan, ks. (McTear, 2002, s. 94, kohta 2.3 ja s. 100–103, kohta 3.3).

Matkapuhelimia varten on nykyään hyvin edistyneitä ns. älykkäitä ohjelmistoagentteja (engl. intelligent software agent)[6] kuten Applen *Siri* ja Googlen *Google Now*. Näille voi esittää hyvinkin monenlaisia kysymyksiä, jotka koskevat maantiedettä, ravintoloita, yleistietoa jne. tai komentaa matkapuhelimen toimintoja. Nämä järjestelmät ovat selvästikin agenttipohjaisia eli ensin arvaavat, mistä on kyse ja sitten lähettävät kyselyn eteen päin asiaan erikoistuneelle agentille. Kumpikin edellä mainituista toimii puheohjauksella.

5.4 Vuorovaikutuksen kieliteknologiaa

Tarvittavia tekniikoita voivat olla seuraavat:

- *Puheentunnistus* joko siinä muodossa että tunnistetaan vakiosanoja tai -fraaseja tai sitten sanoista yksinkertaisella kieliopilla yhdisteltävissä olevia lausekkeita. Näiden täydennyksenä sanojen tai lausekkeiden tunnistamista muun viestin keskeltä (word spotting, concept spotting).

- Yksinkertaistettuja *kielioppeja ja jäsentimiä*. Näitä tarvitaan kuvaamaan kyseisen tehtävän kannalta relevantti osa käyttäjän vuorosanoista.

[6] Katso Wikipediasta tämännimistä artikkelia, jossa viitteitä.

- Aihealuetta koskevaa käsitteistöä, *käsitteiden suhteita ja hierarkiaa* kuvaavia ns. ontologioita, joiden avulla voidaan usein päätellä, mihin keskustelija viittaa ja mitä hän tarkoittaa.

- *Dialogin hallinta*, eli päätteleminen miten käyttäjän syöte muuttaa tilaa ja mitä seuraavaksi pitäisi tehdä. Suunnittelua siitä, minkä kokoinen annos olisi hyödyllisintä sijoittaa seuraavaan repliikkiin.

- Lauseiden ja vastausten *generointi* sanalliseen asuun em. suunnitelmasta.

- *Puhesynteesi* käsitteestä puheeksi (eikä niinkään tekstistä puheeksi) -muotoisena.

- Kerättyjen ja pääteltyjen tietojen varmistamisen ja *tarkistamisen menettelytavat* ja mitä tehdään, kun yritykset tulkita käyttäjän puheenvuoroa epäonnistuvat eli valistuneet arvaukset siitä, mistä suunnilleen saattaisi olla kyse, jne.

5.5 Kysymys–vastaus -järjestelmät

Kysymyksiin vastaava tietokoneohjelma on sukua toisaalta *tiedonhakutehtävälle* ja toisaalta dialogisovelluksille. Varsinaista dialogia niissä ei tarvitse olla. Yhteen kysymykseen riittää yksi koneen löytämä vastaus (tai arvaus vastaukseksi).

Yleensä tehtävä voidaan ajatella esimerkiksi sellaisena, että meillä on käytettävissä runsaasti dokumentteja, joissa on tietyltä alueelta oikeata tietoa. Tältä alueelta järjestelmä pyrkii antamaan vastauksia ja vastaukset perustu-

vat *vain siihen tekstiin*, joka dokumenteissa on, vaikkakin tekstejä voidaan lyhennellä ja leikellä sopivamman vastauksen tuottamiseksi.

Tyypillinen tarve kysymys–vastaus-järjestelmälle voisi olla tietokoneohjelman käyttöopastus. Käyttäjä saattaisi kysyä MIKÄ ON TYYLITIEDOSTO? ja kelpuuttaisi vastaukseksi käyttäjän käsikirjasta sellaisen kohdan, jossa tyylitiedosto määritellään. Vastausjärjestelmäksi ei riitä tiedonhaku eli sellaisten dokumenttien löytäminen, jotka ovat relevantteja asian suhteen. Pitäisi löytää dokumentti, jossa on vastaus esitettyyn kysymykseen ja lisäksi *eristää* vastaus dokumentista tai ainakin osoittaa dokumentista kohta, jossa vastaus on. Esimerkin tapauksessa niistä kohdista, joissa puhutaan tyylitiedostoista, halutaan sellainen, jossa määritellään käsitettä, esim. TYYLITIEDOSTOLLA TARKOITETAAN …, vrt. (McTear, 2002, luku 31, s. 560–582).

Aiemmin mainittujen Google Now ja Applen Siri -järjestelmissä on osana myös kysymys–vastaus-järjestelmän mukaan toimivia agentteja.

Luku 6

Kieliteknologia kielen oppimisessa ja opettamisessa

6.1 Taustaa

Eurooppa on sitoutunut monikielisyyteen ja sadat miljoonat ihmiset osaavat ainakin jossakin määrin muitakin kieliä kuin äidinkieltään. Kenties monikielisyys on ollut kautta aikojen pikemmin sääntö kuin poikkeus. Monikielisyys koskettaa siis valtavan suurta osaa ihmisistä eikä tätä yleensä pidetä ongelmana. Euroopan unionin poliittisen tahdon mukaan ihmisiä tulisi kannustaa pikemminkin *lisäämään kielitaitojaan*. Kielitaidolla on toki eri asteita auttavasta luetun ymmärtämisestä ja ymmärretyksi tulemisesta aina sujuvaan kaksikielisyyteen, jossa näitä kieliä osataan käyttää yhtä luontevasti ja virheettömästi.

Tietokoneavusteinen kielenoppiminen (tai -opettaminen)

eli CALL (Computer Assisted Language Learning) muodostaa varsin laajan kentän, joka voi tulevaisuudessa koskettaa satoja miljoonia ihmisiä. Käytännöiltään ja teoriataustaltaan CALL liittyy *soveltavaan kielitieteeseen*, mutta siinä voidaan käyttää monin tavoin hyväksi *kieliteknologian menetelmiä* ja työkaluja. Kaikki tietokoneavusteinen kielenoppiminen ei suinkaan liity kieliteknologiaan vaan monet sen sovelluksista perustuvat aivan tavanomaiseen tietotekniikkaan.

Tietokoneavusteisessa (tai teknologia-avusteisessa) oppimisessa sovelletaan toisaalta *tekstipohjaista* kieliteknologiaa kuten jäsentimiin perustuvaa ja toisaalta *puheteknologiaa*, mutta lisäksi tarvitaan tietoa myös ihmisen ja koneen välisistä *käyttöliittymistä* ja vuorovaikutuksesta. CALLissa on siis kyse erityisen monitieteisestä alueesta.

6.2 Tavanomainen tietokoneavusteinen kielenopiskelu

Nykyiset tietokoneet käsittelevät helposti *multimediaa*, joka koostuu *tekstistä*, paikallaan olevaa *kuvista* ja *liikkuvasta kuvasta* sekä *äänestä*. Monet tietokoneavusteisen kielenopiskelun ohjelmat hyödyntävätkin juuri näitä ominaisuuksia. Silti voidaan sanoa, että pelkkä multimedia ei tee vieraan kielen opiskelusta erityisen tuloksellista tai mielekästä. Sen avulla *jäljitellään usein vain perinteistä kirjekurssia*, jossa oppilas sai luettavan ja kuunneltavan materiaalin postilähetyksenä ja lähetti vastauksensa lomakkeella rastimalla kysymysten vaihtoehtoja ja täyttämällä

tyhjiin kohtiin sanoja. Videonauhakasetin sijasta ohjelma voi suoraan näyttää oppilaalle liikkuvaa kuvaa tietokoneen ruudulla ja soittaa siihen liittyvää puhetta tietokoneen kaiuttimista tai kuulokkeista. Tietokoneohjelmalla kirjekurssin toteuttaminen on nykyään helpompaa, nopeampaa ja halvempaa kuin postilähetyksinä, muttei välttämättä oppimisen kannalta kovin erilaista.

Vuorovaikutus oppilaan ja opettajana toimivan tietokoneen välillä jää tällaisissa sovelluksissa vähäiseksi. Ellei opetuksessa ole mukana myös ihmisopettajaa, oppilas *ei saa kovinkaan ymmärtävää palautetta* ohjelmalta. Siinä suhteessa perinteinen kirjekurssi, jossa opettaja on mukana kuvioissa, vaikkakin kaukana, voisi olla edelleen parempi kuin pelkkään yksinkertaiseen tietotekniikkaan nojaava sovellus.

Muita tavanomaisia tietokoneavusteisia kielenopiskelumenetelmiä ovat erilaiset *sähköpostin* ja *keskusteluryhmien* (chat) käyttöön perustuvat sovellukset tai tietokonepohjaisen *videoneuvottelutekniikan* hyödyntäminen, joiden avulla saadaan tavanomaista vuorovaikutusta oppilaan ja opettajan välille tai opiskelevan ryhmän kesken. Näissä *ei* yleensä *sovelleta* lainkaan *kieliteknologian* menetelmiä.

6.3 Ymmärtävämpää kielenopiskelua kieliteknologian avulla

Meidän kannaltamme on tärkeää ottaa esille se, että kieliteknologian avulla tietokone voidaan saada enenevässä määrin *ymmärtämään* ainakin kielen rakennetta ja sään-

nönmukaisuuksia. Tietokone pystyy varastoimaan ja käsittelemään toisaalta *suuria tietomääriä* kuten sanakirjoja ja toisaalta se voi *nopeasti käsitellä* tietoja monimutkaistenkin sääntöjen mukaan. Koneen rajoitukset eivät nykyään juuri estä toteuttamasta ohjelmia, jos vain tiedetään tarkasti, mitä ohjelman halutaan tekevän. Kielenoppimisen kieliteknologian haasteena olisi siis *yhdistää* kieliteknologian tietämys ja menetelmät hyödylliseksi osaksi kielenopiskeluohjelmia.

Kieliteknologian ja CALLin yhdistämisessäkään *ei* ole kyse siitä, että sen avulla pyrittäisiin ensisijaisesti korvaamaan ihmisopettaja tietokoneella, eikä tämä olisi nykyisen tiedon ja teorian puitteissa edes mielekäs tai lähiaikoina saavutettavissa oleva tavoite. Sen sijaan kieliteknologian avulla voidaan tehdä opettajalle ja oppilaille *apuvälineitä*, jotka helpottavat tai tehostavat työtä.

Esimerkkinä kieliteknologian mahdollisuuksista voimme ajatella, että järjestelmä voisi voida antaa rakentavaa ja oikein kohdennettua palautetta, ei ainoastaan ”oikein” tai ”väärin”. Järjestelmän tulisi siis tunnistaa *missä suhteessa ja miten* oppilaan vastaus on väärä ja *kohdentaa palaute* sen mukaisesti. Lisäksi ohjelma voi seurata, mitä sääntöjä opiskelija jo hallitsee ja vastaavasti, missä hän tarvitsee vielä lisää harjoitusta. Tällainen ohjelma voisi auttaa *yksilöllisesti* kutakin opiskelijaa kohdistamaan työskentelynsä tarkoituksenmukaisesti.

Toisaalta siis järjestelmän pitäisi tunnistaa oppilaan *osin oikeista* mutta *osin virheellisistä* tuotoksista, mitä oppilas on yrittänyt ilmaista ja neuvoa häntä tilanteen ja tarpeen mukaan valikoivasti. Jos esim. ulkomaalainen opiskellessaan suomea tuottaisi taivutusmuodon:

kun tarkoitus olisi tuottaa KAUPOISSA, voisi ohjelma havaita, että oppilas osannee inessiivin ja astevaihtelun, mutta A-loppuisen substantiivin vartalonlopun vokaalivaihtelu tarvitsee vielä selitystä ja harjoitusta. Sopivilla kieliteknologian malleilla tällaista analyysiä voidaan tehdä melko yleisesti ilman, että oppimateriaalin laatijan pitäisi yrittää luetella etukäteen mahdollisia virhetuotoksia selityksineen.

Varmaan perinteisessäkin kielenopetuksessa, tapahtuipa se luokassa, kirjekurssina tai tietokoneohjelman tukemana, *oppilaasta ja hänen taidoistaan hahmotetaan jonkinlainen malli*. Kieliteknologian avustama kielenoppiminen tarjoaa lisää mahdollisuuksia mallintamiselle. Se lisää myös mahdollisuuksia tehdä havaintoja siitä, miten opiskelija edistyy tavoitteissaan oppia kieliopin sääntöjä.

Kieliteknologian menetelmillä voidaan tunnistaa oppilaan tekemiä kielivirheitä. Oikeinkirjoitusvirheiden tunnistaminen ja korjausehdotusten teko ovat läheistä sukua kirjoittajan apuvälineinä tunnetuille työkaluille. Kieliopillisuuden tarkistus (grammar checking, grammatikkontroll) on myös käyttökelpoinen väline aloittelevalle kielenopiskelijalle. On raportoitu, että oppilaat mieluusti kokeilevat tällaisella ohjelmalla tekstiään ennen, kuin antavat sen opettajalle arvosteltavaksi ja tarkastettavaksi. Tietokoneohjelma on huomaavainen ja siltä saa nopeasti palautteen, josta usein on ainakin jotakin hyötyä. Opettajaa ohjelma ei tietenkään korvaa, mutta useinkin parantaa opiskelun tehoa.

Vieraskielistä tekstiä lukemalla oppii kieltä. Lukemista voi tukea ja sen avulla oppimista voi tehostaa kielitek-

nologisilla apuvälineillä. Vieraan kielen sanaston puutetta voidaan paikata Xeroxin tutkimuslaitoksessa kehitetty *Locolex*-hankkeen kaltaisilla välineillä. Siinä yhdistetään kolme komponenttia:

- *kaksikielinen sanakirja*, joka antaa tekstissä esiintyville sanoille käännöksiä lukijan äidinkielelle,
- *morfologinen jäsennin*, joka palauttaa tekstissä olevan saneen perusmuotoonsa, jotta haku sanakirjasta onnistuisi sekä
- sanan *alamerkityksen päättelyohjelma*, joka laskee saneen ympäristön perusteella todennäköisyyksiä hakusanan eri alamerkityksille kyseisessä kontekstissa, minkä perusteella vaihtoehtoisista käännöksistä voidaan tarjota ensimmäisenä se, joka on luultavimmin oikea.

Käyttäjä lukee tekstiä ilman sen kummempia pysähdyksiä niin kauan kuin ymmärtää. Usein jopa neuvotaan käyttäjää lykkäämään haluaan käyttää sanakirjaa siinä toivossa, että asiayhteys paljastaa merkityksen. Vasta koko virkkeen lukemisen ja virkkeen pääasiallisen rakenteen hahmottamisen jälkeen olisi siis hyvä palata mahdollisiin jäljelle jääviin ongelmasaneisiin. Kielenopiskelija saa ohjelmalta käännökset näkyvilleen napsauttamalla hiirtä tuntemattomaksi jääneen saneen kohdalla. Tampereen yliopiston koordinoima iEye-niminen hanke tutki tämän viemistä vielä askelta pidemmälle yhdistämällä yllä mainitut tekniikat silmänliikkeiden tunnistimeen, jolloin katseen

pysähtyminen tiettyyn saneeseen havaitaan ja se laukaisee käännösten tarjoamisen.[1]

Locolexin eräs jatkohanke, Glosser-projekti tuotti hollanninkielisille käyttäjille välineitä ranskankielisten tekstien lukemisen helpottamiseksi.[2]

6.4 Oppimateriaalien tuottaminen kieliteknologian avulla

Kielitieteellistä tietojenkäsittelyä on kauan käytetty kielentutkimuksen apuna esimerkkien etsinnässä ja sanojen tyypillisten käyttötapojen tunnistamisessa. Tyypillisiä perinteisiä apuvälineitä ovat olleet *taajuussanakirjat*, joiden avulla voidaan paitsi tutkia kieltä, myös valita oppimateriaaliin keskeisintä ja tarpeellisinta sanastoa. Näistä tietokoneella tuotettavista listoista tai hakujärjestelmistä tulee käyttökelpoisempia, jos *morfologinen jäsennin* on käytettävissä ja vieläkin käyttökelpoisempia, jos lisäksi käytettävissä on tulkintoja yksiselitteistävä ohjelma ns. *sanaluokkajäsennin* (engl. part of speech tagger, morphological disambiguator). Näiden välineiden avulla voidaan helpommin löytää lekseemin esiintymät vaikka sana taipuisi hyvinkin erilaisiin muotoihin.

Joskus ei ole käytettävissä sanakirjaa, mutta kuitenkin tietokoneen muodossa olevia tekstejä. Silloin outojen sanojen merkityksistä voi saada selkoa ns. konkordanssien avulla. *Konkordanssi* on jo kauan (siis useita satoja vuo-

[1] Ks. esim. http://www.cs.uta.fi/research/hci/ieye/

[2] Ks. mm. http://www.let.rug.nl/glosser/

sia) tunnettu tapa listata jonkin tekstin tai tekstikokoelman kaikkien saneiden esiintymät aakkosjärjestyksessä esiintymisyhteytensä kera.

```
1363:          Ja otettuaan  ryypyn jatkoi kapteeni, että
2441: minen ja he kumosivat  ryypyn kaikki.
2896:           – Minä otan  ryypyn!
710:     Mutta kun isä otti  ryypyn, niin otti Anttikin.
```

Usein saman sananmuodon esiintymät kohdistetaan allekkain siten, että aakkostettu sananmuoto on samalla kohdalla sivua kuten yllä (ns. KWIC eli keyword in context -konkordanssi). Joskus konkordansseja työstetään morfologisen jäsentimen avulla sellaiseksi, että saman hakusanan eri sananmuodot aakkostaan yhteen ja varustetaan hakusanalla eli lemmalla. Tätä kutsutaan lemmaukseksi (engl. lemmatization).

Tiedon jalostamiseksi kutsutaan sitä, kun laajemmista tietomassoista eristetään tarkoitukseen hyödyllisiä osia, joita tarvittaessa vielä muokataan tarkoitukseen sopivammiksi. Tietomassoina voivat toimia Internetistä *hakukoneella löydettävissä olevat tekstit* ja valikointi voidaan tehdä kieliteknologisin kriteerein ja välinein. Prosessissa voidaan käyttää tiedonhausta tuttuja välineitä: mm. saneiden palauttamista perusmuotoonsa, tekstin tiivistämistä, indeksointia, termien tunnistamista.

Erityinen tiedon jalostamisen muoto on *BootCaT* -niminen ohjelma, jolla Internetistä eristetään tiettyyn asiakokonaisuuteen liittyvän kielen opiskelua varten *valikoituja korpuksia*, ks. Baroni & Bernardini (2004).[3] BootCaT käyttää hyväkseen verkon tiedonhakuohjelmia (kuten Yahoota) ja lähtee etsimään tiettyjen siementermien mukaan si-

[3]Ks. myös http://bootcat.sslmit.unibo.it/

vuja, joilla nuo sanat esiintyvät löytääkseen lisää näihin liittyviä sanoja. Muutaman uusintakierroksen jälkeen päädytään korpukseen, joka koostuu kyseisen teeman teksteistä. Näin saatua korpusta voidaan käyttää erilaisiin tarkoituksiin, mm. erityisalan termien keräämiseen, mutta myös kielenopiskelun tarkoituksiin, kuten Reetta Vuokon pro gradu -työssä (2009). Siinä BootCaTia käytettiin etsimään *vaikeusasteeltaan ja sanastoltaan sopivia tekstejä* japanin kielen opiskelua varten. Sanaston helppoutta voidaan helposti arvioida, mutta tekstin muunlainen helppous liittyy luettavuuden arvioimisen tehtävään. Kyseisessä työssä luettavuutta arvioitiin automaattisella japanin kielen SYNTAKTISELLA JÄSENTIMELLÄ. Rakenteen helppoutta voidaan arvioida jäsentimen tuottaman rakennekuvauksen perusteella. Pelkän virkkeiden pituuden sijasta voidaan käyttää myös sitä, millaisia *kieliopillisia muotoja* sanoista käytetään sekä millaisia *syntaktisia konstruktioita* virkkeissä on.

6.5 Kaksikieliset korpukset

Kaksikieliset tekstit ovat kielenoppijalle usein hyödyllisempiä kuin yksikieliset. Ennen tietokoneiden aikaa julkaistiin kirjoja, joissa kunkin aukeaman toisella sivulla oli alkukielinen teksti ja toisella sivulla sen käännös. Tietokoneiden aikana vastaavaa kaksikielisen materiaalin tahdistamista kutsutaan *kohdistamiseksi* (englanniksi alignment). Teksti kohdistetaan kuitenkin sivujen sijasta yleensä kappaleittain tai virkkeittäin ja joskus vieläkin pienempinä yksiköinä eli *käännösvastineina*.

Kohdistetut kaksikieliset korpukset olisivat kullan ar-

voisia välineitä kielenoppijalle, kunhan on käytettävissä niitä hyödyntäviä sovelluksia. Itse asiassa verkkosivujen kuvauskieli HTML antaa suoraan välineet esittää kohdistettua kaksikielistä tekstiä lukijan kannalta hyödyllisessä muodossa, kunhan kohdistetut tekstit yhdistetään mekaanisesti siten, että alkukielinen teksti näkyy oletuksena ruudussa ja käännös tulee näkyviin vain, jos hiirikohdistin viedään sellaisen virkkeen kohdalle, johon kaivataan selvennystä.

6.6 Puheteknologian mahdollisuudet kielenoppimisessa

Opittavaan kieleen liittyy paitsi sen kirjoitettu muoto, myös sen ääntämys. Puhesynteesin avulla mikä tahansa teksti saadaan kuultavaan muotoon. Ennalta tiedetyt ja kiinteät sanottavat voidaan toki äänittää ilman puheteknologiaakin, mutta vuorovaikutteisissa kielenoppimisessa tarvitaan helposti myös uusien, esim. käyttäjän itsensä tuottamien tai vaikkapa aineistosta tai verkosta löydettyjen tekstinpätkien ääneen puhumista.

Voidaan ajatella, että kielenoppijalle on hyötyä siitä, että opittavaan kieleen liittyy *mielikuva sen äänneasusta* eikä vain kirjoitetuista muodoista. Ainakin puhesynteesistä on selvää hyötyä kielenoppimisessa sellaisissa olosuhteissa, joissa voi hyvin kuunnella, mutta suurta näyttöruutua ei ole käytettävissä.

Turun yliopistossa on kehitetty niin kutsuttu *vokaalipeli*, jonka avulla oppilas voi harjoittaa sellaisten vieraan

kielen vokaalien ääntämystä, joita hänen äidinkielessään ei samanlaisina ole Paganus et al (2006). Vokaalipeli perustuu *äänen reaaliaikaiseen analyysiin*, jossa vokaalista eristetään spektrin kautta sen päätaajuuksia eli formantteja. Formanttien avulla vokaali, esimerkiksi ruotsin kielen 'u' voidaan sijoittaa kaksiulotteiseen karttaan, jossa myös tavoiteltava ääntämys on merkittynä (suomen kielen 'u':n ja 'y':n välimaille). Oppilas näkee kuviosta, kuinka lähelle tavoiteltua vokaalia hän kulloinkin osuu ja milloin vihdoin onnistuu kokonaan.

Kun oppilas harjoittaa vieraan kielen ääntämystä, voidaan *puheentunnistusta* käyttää ohjaamaan ääntämystä tavoiteltuun suuntaan. Tunnistin voisi esimerkiksi olla harjaannutettu tunnistamaan oman kielen mukaisille puutteellisille ääntämyksille ja oikeille ääntämyksille rajatulla sanastolla (esim. englannin kielen 'thirteen' lausuttuna niin, että paino on lopputavulla kuten pitää tai virheellisesti sanan ensitavulla). Oppilaan harjoitellessa tunnistin sitten luokittelisi tuotoksen joko oikeaksi tai ennakoiduksi vääräksi ja antaisi asianmukaista palautetta ja ohjeita ääntämyksen korjaamiseksi.

Vokaalipelin tapaan myös äänisignaalin muutakin kuvantamista esimerkiksi spektrogrammeina voidaan ajatella käytettäväksi ääntämisen harjoittelussa. Erityinen ja haasteellinen tehtävä on, kun *kuuro opettelee ääntämään*. Kuuro voisi oletettavasti saada tällaisen kuvantamisen avulla paljon arvokasta hyödyllistä palautetta, jota ihmisopettajankin on vaikea antaa.

6.7 Johtopäätöksiä

Kielenoppimisella ja kieliteknologialla pitäisi olla paljon enemmän yhteistä kuin käytännössä on vielä toteutettu. Ollakseen hyödyllistä kieliteknologisen tietokoneavusteisen kielenopetuksen ei tarvitse olla täydellistä eikä edes yhtä hyvää kuin ihmisopettajan tarjoama opetus. Tietokoneet ovat halpoja, uupumattomia ja hienotunteisia ja niitä on laajalti saatavissa. Riittää siis, että CALL-menetelmien käyttämisestä on hyötyä siten, että niiden käyttäminen tuottaa parempia tuloksia kuin se, että niitä ei käytetä.

Joka tapauksessa tällä alueella lienee valtava potentiaali, koska se koskettaa hyvin suurta osaa maailman väestöstä.

Luku 7

Kielen kääntämisen apuvälineet ja automaattinen kielen kääntäminen

7.1 Monikielisyys ja kääntämisen tarve

Maailmassa puhutaan tätä nykyä muutamia tuhansia kieliä ja Euroopassakin huomattavaa määrää. Aiemmin on jo todettu, että Euroopan unioni on poliittisesti sitoutunut tukemaan alueellaan käytettyjä kieliä, myös vähemmistöjen kieliä. Virallisen aseman omaavia kieliä EU:ssa on tätä nykyä 23 ja kaikkiaan Euroopassa voidaan arvioida olevan satakunta kieltä, jolla on jonkinasteinen ainakin kan-

sallinen status.[1] Vaikka politiikkana onkin tukea sitä, että yksilöt osaavat muitakin kieliä kuin äidinkieltään, ei pelkkä kielten opiskelu ratkaise monien rinnakkaisten kielten tuottamia ongelmia. Paikallisten kielten käytön ja kansainvälisen kanssakäymisen välille syntyy tietty ristiriita, sillä:

- paikallista, varsinkaan pientä kieltä ei ymmärretä muissa maissa juuri lainkaan,
- paikallinen siirtyminen valtakielen, esimerkiksi englannin, käyttämiseen, uhkaa oman kielen ja kulttuurin elinkelpoisuutta,
- eivätkä ihmiset ole kovin tarkkoja ja taitavia ilmaisemaan ajatuksiaan vieraalla kielellä, esim. englanniksi.

Kauppa ja kanssakäyminen hyötyvät kuitenkin valtakielien käytön lisääntymisestä. Monet suuret monikansalliset firmat edistävät valtakielten asemaa siirtymällä koko organisaatiossaan esimerkiksi englannin kielen käyttämiseen. Firmalle olisi ongelma, jos osa sen dokumentaatiosta, asiakirjoista ja suunnitelmista olisi vain sellaisella kielellä, jota vain harvat firman organisaatiossa osaavat.

Kielten rinnakkaiseloon liittyy tehtäviä, joista monet sivuavat kieliteknologiaa:

- *vieraan kielen oppiminen* ja opettaminen (josta oli puhetta edellisessä luvussa),

[1] Ks. esim. http://en.wikipedia.org/wiki/Official_languages_of_the_European_Union

- tekstien, kuten ohjeiden, teknisten käsikirjojen, uutisten, kaunokirjallisuuden ym. *kääntäminen* vieraalle kielelle, jota tämä luku erityisesti käsittelee,
- *tulkkaus* eli puhutun kielen kääntäminen toiselle kielelle,
- paikalliskielisten *uudissanojen ja termien* keksiminen uusille tuotteille ja ilmiöille,
- tuotteiden sovittaminen *paikallisilla kielillä toimiviksi* (eli ns. lokalisointi),
- tuotteiden valmistaminen sellaiseksi, että ne voidaan tarjota *useille kielille sovitettuna* (eli ns. internationalisointi) sekä
- *tiedon hakeminen monikielisistä dokumenteista*.

7.2 Kielen kääntämisen vaativuus

Ilmauksia tai tekstejä kieleltä toiselle käännettäessä tulee esille monia niitä kielten järjestelmiin liittyviä seikkoja, joista johdantoluvussa oli puhe. *Sananmuotojen moniselitteisyyden* vuoksi on joskus vaikeuksia tunnistaa, mikä lekseemi moniselitteisissä tapauksissa on kyseessä. Aiemmin todettiin myös, etteivät kielten syntaktiset järjestelmät koodaa yksiselitteisesti kielen merkityksiä, vaan usein tekstin ilmauksille jää useampia *vaihtoehtoisia tulkintoja*. Kääntämisen vaikeus riippuu tältä osin siitä, *kuinka samanrakenteinen* kohdekieli on lähtökielen kanssa. Samanrakenteisten kielten kesken ei haittaa niin paljon, vaikka

rakenteen tulkitsisi väärin, kunhan sen kääntää samanrakenteisena. Väärin hahmotettu voi silloin kuitenkin kääntyä oikein tulkittavaksi, koska kohdekielessä voi vallita sama rakenteellinen moniselitteisyys.

Kielen yleisten ominaisuuksien lisäksi käännettäessä tulee muitakin ongelmia selvästi näkyville. Lekseemeillä on *alamerkityksiä* ja eri alamerkityksillä voi olla erilainen käännösvastine. Esim. englannin kielen verbi PLAY kääntyy suomeksi mm. PELATA, LEIKKIÄ, NÄYTELLÄ tai SOITTAA sen mukaan, millaisesta toiminnasta on kyse. Ei siis riitä vain se, että löydetään oikea lekseemi, vaan pitäisi myös tunnistaa lekseemin oikea alamerkitys, esim.:

- engl. BROTHER tai suomen VELI: japaniksi nuorempi veli on OTOOTO ja vanhempi veli ONIISAN; mandariinikiinaksi vastaavasti vanhempi veli on GEGE ja nuorempi DIDI,
- engl WALL: saksaksi sisällä oleva seinä on WAND ja ulkona oleva seinä tai muuri on MAUER,
- englannin THEY tai suomen HE: ranskaksi naispuolisesta ELLES ja miespuolisesta ILS ja
- saksan BERG: englanniksi joko HILL tai MOUNTAIN sen mukaan kuinka isosta kukkulasta, mäestä tai vuoresta on kyse.

Joskus kääntäminen siis edellyttää *alamerkityksen valintaan tarvittavaa tietoa*, joka ehkä on käännettävässä tekstissä ilmaistuna jollakin tavoin ehkä epäsuorasti. Merkitystä ei kuitenkaan aina voi päätellä itse tekstistä. Jos vaikka Raamatussa olisi puhuttu jonkun veljestä, pitäisi

tiettyihin kieliin käännettäessä kääntäjän tietää, oliko kyse nuoremmasta vai vanhemmasta veljestä, eikä tätä välttämättä ole Raamatussa kerrottu.

Kielet jäsentävät todellisuutta eri tavoin. Ongelma ei vain rajoitu siihen, että lekseemeillä on alamerkityksiä, jotka kääntyvät eri tavalla toiseen kieleen. Jos tarkastelemme useampaa kuin kahta kieltä, tilanne voi entisestään hankaloitua. Alamerkitykset eivät nimittäin ole *universaaleja*, siten että kaikki kielet voitaisiin koostaa siedettävästä määrästä yhteisiä alamerkityksiä. Useampien kielten kesken nämä alamerkitykset pilkkoutuvat yhä pienemmiksi, eikä ehkä voida edes vakavasti kuvitella rakennettavan kaikkia maailman tuhansia kieliä kattavaa *yhteistä alamerkitysten järjestelmää*. Vaikka kuvittelisimme, että meillä olisi sellainen, ei se ratkaisisi ongelmia. Jos sanoilla olisi sadoittain tällaisia hienojakoisia alamerkityksiä, olisi meillä edessä vain entistä valtavampi moniselitteisyyden ratkomistehtävä.

Kääntäminen tulee erityisen vaativaksi silloin, kun *yhteiskuntaan ja kulttuuriin liittyvät rakenteet ja käsitteet ovat erilaisia*. Esim. koulujärjestelmät ja sosiaaliturva voivat jäsentyä eri kulttuureissa eri tavoilla — niinkin, että kohdekielessä ei ole vakiintuneita ilmauksia lähtökielessä tavanomaisille asioille. Kaikille asioille ei siis välttämättä ole kunnollista käännettyä ilmausta, vaikka kuinka yrittäisimme kääntää kuinka huolellisesti ja tarkkaan.

Eri kulttuureissa on myös erilaisia tapoja lähestyä asioita. Toisissa kulttuureissa voidaan mennä suoraan asiaan, toisissa edellytetään tiettyjä kohteliaisuuksia tai muodollisuuksia. Se vuoksi joskus joudutaan *kääntämään tulostekstiin sellaista, mistä alkutekstissä ei ole sanaakaan*

tai päinvastoin jättämään käännettäessä pois alkutekstin osia, joita ei ole tapana kohdekielessä käyttää.

7.3 Kääntäjien kieliteknologisia apuvälineitä

Kääntäjän työtä voidaan avustaa joko välineillä, jotka parantavat käännöksen laatua tai sitten välineillä, jotka nopeuttavat ja tehostavat kääntämistä. Riippuu tehtävästä, kuinka tärkeää *käännöksen laatu* on. Yleensä työn tulokselta vaaditaan tietty korkeahko laatutaso, mutta se pitäisi tehdä sen puitteissa mahdollisimman *taloudellisesti*. Joissakin tehtävissä saattaa kuitenkin olla kaikkein tärkeintä saada käännös *nopeasti* valmiiksi, sillä esimerkiksi huomisen päivän sääennustus ei vuorokauden viiveen jälkeen enää ole paljonkaan arvoinen.

7.3.1 Terminologian hallinta

Terminologia on olennainen osa käännöstyötä. Laadukkaan käännöksen aikaansaamiseksi kääntäjä yleensä perehtyy lähdetekstiin määritelläkseen siinä käytetyt termit. Tässä voidaan käyttää apuna kieliteknologisia työkaluja, jotka esimerkiksi tunnistavat tekstin *substantiivilausekkeet* tiettyjen kriteerien perusteella. Tällainen mekaanisesti tehty lista voi vielä olla järjestettynä kieliteknologian välineillä arvioidun *termimäisyyden* perusteella, jolloin listan yläpäässä on enimmäkseen todellisia ja käytettyjä termejä. Listan hännillä ovat vastaavasti epävarmimmat termiehdokkaat.

Termimäisyyden arvioimiseksi on hyödyllistä kerätä

tilastoja. Johdonmukaisten tilastojen laskemiseksi voidaan yleensä käyttää *morfologista jäsennintä*, joka palauttaa eri taivutusmuodot samaan perusmuotoon. Tilastot, termikandidaattien morfologiset (esim. johtamista koskevat) ominaisuudet sekä esiintymisyhteydet voivat toimia termien valinnan hyödyllisinä kriteereinä. *Syntaktisen jäsentimen* avulla kriteereiksi saadaan lisäksi esim. esiintyminen tiettynä lauseenjäsenenä (esim. objektina ylipäätänsä) tai tiettyjen verbien argumentteina. *Sanojen yhdistelmät* voivat myös olla termejä, joten tilastoja lasketaan myös niistä. Kriteereistä lasketun kaavan mukaan voidaan termikandidaatit järjestää yleensä varsin hyödylliseen järjestykseen, jonka mukaan voidaan valita sopivaksi katsottu määrä termejä paremmasta päästä.

Kun kääntäjällä on joko kieliteknologisin menetelmin tai muutoin tehty termiluettelo, pitäisi toki voida käyttää sopivia apuvälineitä sen tarkistamiseksi, että käännetyssä tekstissä tätä *termistöä on käytetty johdonmukaisesti*.

Kaupallisestikin on saatavissa eräitä ohjelmia termikandidaattien poimimiseksi ja termitietokantojen ylläpitämiseksi. Täysin tyydyttäviä tuotteita ei kuitenkaan vielä ole saatavilla, vaan tarvitaan lisää tutkimusta ja kehitystyötä. Näistä asioista puhutaan enemmän käännösteknologian kursseilla.

7.3.2 Sähköiset sanakirjat

Kääntäjä tarvitsee toki sanakirjoja. *Digitaaliset sähköiset sanakirjat* ovat helppokäyttöisiä siinä ympäristössä, missä nykyajan kääntäjä muutenkin tekee työtänsä eli tietokoneella. Sanakirjojen sähköiset versiot mahtuvat kenties

mukavammin kääntäjän työpöydälle kuin paperille painetut. Sanojen etsiminen sähköisistä sanakirjoista on yleensä vähintään yhtä helppoa kuin paperisista, parhaimmillaan osoittamalla vain hiirellä tekstin sanetta.

Sähköinen sanakirja on hyödyllinen jo sellaisenaan, mutta sähköiseen versioon voidaan yhdistää myös erilaista kieliteknologiaa parantamaan niiden käytettävyyttä. Ensinnäkin *morfologista jäsennintä* hyödyntävälle sanakirjalle voisi osoittaa sananmuodon sellaisena, kun se alkutekstissä on ilman, että käyttäjän tarvitsee sitä ensin palauttaa perusmuotoonsa haun mahdollistamiseksi. Perusmuotoon palauttavaa sanakirjaa käytettäessä haku siis helpottuu, mutta samaa teknologiaa hyväksi käyttäen sanakirjaohjelma voi vielä antaa käännöksen taivutusmuodossa, joka parhaiten vastaa lähtökielen puolella ollutta muotoa, jolloin sen kopioiminen käännökseen on helpompaa.[2]

Sanakirja tarjoaa yleensä käännöksiä lekseemin (eli hakusanan) eri alamerkityksille siinä järjestyksessä, jonka sanakirjan laatija on valinnut: joissakin se on se ikäjärjestys eli se järjestys, jossa alamerkityksistä on historiallisesti todennettuja käyttöesimerkkejä, ja toisissa puolestaan alamerkitysten yleisyysjärjestys. Sähköinen sanakirja voi kuitenkin helposti vaihdella alamerkitysten järjestystä. Ohjelma voi nimittäin arvioida eri alamerkitysten todennäköisyyksiä ympäröivän tekstin perusteella ja lajitella sen jälkeen alamerkitykset siten, että *todennäköisimmät tulkinnat tulevat listan kärkeen* (vrt. Locolex-hanke, josta kielenoppimisen välineiden yhteydessä oli puhetta).

[2]Eräs tällainen taivutuksen huomioiva sanakirja:
http://www.ling.helsinki.fi/kieliteknologia/tutkimus/finnwordnet/demot.shtml

7.4 Automaattiseen kielen kääntämiseen liittyvät kieliteknologiset sovellukset

Kielen kääntämiseen liittyy eri asteisia sovelluksia alkaen yksinkertaisista apuvälineistä aina täysin automaattiseen konekäännökseen. Näihin kaikkiin on syytä suhtautua kuten mihin tahansa työkaluihin eli jos niiden käyttäminen on kannattavampaa kuin käyttämättä jättäminen, niin työkalu on hyödyllinen. Ei puusepälläkään ole täydellistä sahaa, jolla voisi tehdä sileitä pintoja, eikä höylää, jolla voisi kätevästi katkaista lautoja.

7.4.1 Konekäännöksen taustaa

Ensimmäinen varsinainen tietokone valmistui vuonna 1945 ja valjastettiin tiettävästi laskemaan tykin ammusten ballistisia ratoja. Kieleen liittyvät tehtävät olivat kuitenkin jo varhain mukana tietotekniikan kehityksessä. Releohjattuja tietokoneiden edeltäjiä käytettiin toisen maailmansodan aikana Britanniassa saksalaisten Enigma-salakirjoitusten koodin murtamiseen (mikä tiettävästi johti saksalaisten sukellusvenesodan epäonnistumiseen). Tietokoneiden alkuaikana, vuonna 1950 tehtiin jo suunnitelmia tietokoneen avulla tapahtuvaksi *automaattiseksi kielenkääntämiseksi*. Siivittäjänä oli tuolloinen kylmä sota. Amerikkalaiset olivat (ja ovat edelleen) kielitaidottomia ja haluttomia opiskelemaan vieraita kieliä. Silti heidän tiedustelupalvelullaan oli tarvis tietää, mitä kilpailevassa supervallassa Neuvostoliitossa oli tekeillä, minkä vuoksi tilausta konekäännökselle todella oli.

1950-luvun aikana automaattista kääntämistä tutkittiin enenevässä määrin ja 1960-luvun alkupuolen ajan siihen panostettiin hyvin runsaasti varoja. Kun käännösjärjestelmät eivät kuitenkaan näyttäneet valmistuvan lyhyellä tähtäyksellä, katkaistiin rahoitus *Automatic Language Processing Advisory Committee* -ryhmän selvityksen (ALPAC, 1966) perusteella.[3]

Vaikka tuota konekääntämisen alkuaikaa yleisesti pidetään epäonnistuneena, eräs menestyneimmistä käännösjärjestelmistä on peräisin juuri tuon ajanjakson työstä, nimittäin Systran, jota esimerkiksi Euroopan unioni varsin laajasti käyttää. Unionin käännöstarve onkin valtava ja kääntäminen onkin EU:n hallinnon suurin yksittäinen menoerä: yli tuhat kääntäjää on komission ja parlamentin palveluksessa.

7.4.2 Konekäännöksen tavoitteita

Automaattinen kielenkääntö onkin eräs nimenomainen sovelllus, jolle on ollut hyvin vahva tilaus. On ilmeistä, että automaattiselle (varsinkin korkealaatuiselle) kielenkäännölle on ilmeinen ja loputon tarve. Siksi onkin harmillista, että juuri tämä tehtävä on paljon vaikeampi, kuin mitä yleisesti ymmärretään.

Kielenkääntöohjelmille voidaan kuitenkin asettaa eritasoisia tavoitteita:

1. Selvittää, *onko dokumentti kiinnostava* ja kannattai-

[3] ALPAC-komitean mietintö on verkossakin luettavissa, ks. linkki kirjallisuusluettelossa. Raportti on itse asiassa varsin valistunutta ja maltillista tekstiä, eikä ehkä niin tuomitsevaa, kuin mitä kirjallisuudessa on tapana väittää.

siko se kenties antaa kääntäjälle käännettäväksi tarkempaa perehtymistä varten. Tähän riittää melko vaatimaton käännöksen laatu. Vaikka huonosta käännöksestä ei saakaan selvää *mitä* tekstissä sanotaan, siitä voi hyvinkin selvitä *mistä* aihepiiristä tekstissä puhutaan. Tällaisesta käy esimerkiksi Altavistan ja Googlen palvelujen yhteydessä olevat käännökset. Kokeile vaikka kääntää joitakin saksankielisiä verkkosivuja englanniksi, ks. osoite: `http://www.altavista.com/` ja vastaavasti `http://www.google.com`.

2. Tuottaa *raakakäännös*, jonka ihmiskääntäjä *tarkistaa ja korjailee* lopulliseksi versioksi. Raakakäännöksen pitää olla aika hyvä, jotta sen korjailu olisi joutuisampaa kuin tyhjästä aloittaminen. Tällaisesta järjestelmästä on esimerkkinä Kielikone Oy:n TranSmart, joka kääntää asiatekstiä suomesta englanniksi. Demoversio (jonka käyttö tosin ei enää ehkä ole ilmaista) on osoitteessa `http://www.kielikone.fi/` tai ehkä Helsingin yliopiston ns. Nelli-portaalin kautta saatavilla (josta löytyy runsaasti erilaisia sanakirjojakin). Syötä ohjelmalle jostakin raportista kotoisin olevaa asiatekstiä, jotta voit arvioida ohjelman hyödyllisyyttä.

3. Tuottaa lopullinen käännös, joka menee käyttöön, ja jota ei enää erikseen tarkisteta tai korjailla. Tällainen on nykyisellä tekniikalla mahdollista vain suppeilla alueilla kuten säänennustuksessa tai käytettäessä kontrolloitua, tietylle sovellusalueelle räätälöityä typistekieltä.

7.4.3 Käännösmuisti

Eräs laajalti käytetty käännösohjelmien tekniikka on *käännösmuisti*, joka perustuu siihen, että samanlaiset lauseet tai jaksot käännetään toistamiseen esiintyessään samalla tavalla. Ensimmäistä kertaa tietynlajista tekstiä käännettäessä käännösmuistista ei aluksi ole paljonkaan apua, mutta pidemmälle edettäessä alkaa tulla enemmän sellaisia virkkeitä, jotka on jo kertaalleen aiemmin käännetty. Eniten käännösmuistista on hyötyä, kun dokumentista käännetään *uusia versioita*, jotka ovat muuttuneet aiempaan verrattuna vain vähän.

Karkeimmat käännösmuistiin perustuvat sovellukset eivät juuri tunnista kielen rakenteita, vaan joko edellyttävät täsmälleen samoja käännettäviä jaksoja tai soveltavat yleisiä merkkijonojen summittaisen samanlaisuuden kriteerejä etsiessään aiempaa mallia käännökselle. Pidemmälle viety kieliteknologiaa hyödyntävä tekniikka pystyisi soveltamaan *aiemmin käännettyä mallia* eri aikamuotoon ja muutenkin vaihteleviin käyttöihin, esim. vaikka verbin objektina olisi toisenlainen samantapainen kohde taikka henkilön nimen sijasta olisi pronomini.

7.5 Automaattiseen kielen kääntämiseen liittyvät kieliteknologiset menetelmät

Aikanaan haaveiltiin ns. *interlinguasta eli universaalikielestä*, jota käytettäisiin kääntämisen välivaiheena. Sellaisen kanssa voitaisiin suurelta määrältä n kieliä käännetään *miltä tahansa kieleltä mille tahansa* toiselle näistä kielistä

laatimalla $2 \times n$ käännintä eli käännin jokaiselta kieleltä Interlingualle ja takaisin. Interlinguan määrittelemistä rasittavat edellä todetut seikat alamerkitysten yhteismitattomuudesta ja yleensäkin siitä, että kielet jäsentävät maailmaa kukin vähän eri tavalla. Interlingua ei tiettävästi tällä hetkellä ole kovinkaan suuren kiinnostuksen kohteena.

7.5.1 Sääntöpohjainen kääntäminen

Yleisin *sääntöihin perustuvan automaattisen kielenkääntämisen* viitekehys lienee ns. *transfer-menetelmä*, joka poikkeaa interlinguan käyttämisestä siten, että kullekin kielelle laaditaan kyllä erikseen jäsennin ja generoija, mutta kunkin kielen omilla ehdoilla. Jäsentämisen ja tuottamisen pohjana ovat siten sellaiset jaotukset, jotka ovat kielelle ominaisia ja mielekkäitä. Näiden jäsentimien ja generoijien laatimisen katsotaan olevan suuritöistä ja vaativaa verrattuna kokonaisuuteen.

Transfer-mallissa laaditaan kullekin tarvittavalle kieliparille erityinen ohjelma, joka paneutuu kyseisen kieliparin välisiin eroihin, mutta käyttää surutta hyväksi kieliparin välisiä samankaltaisuuksia. Transfer-moduuli muuntaa lähtökielen jäsennysrakenteen sellaiseksi rakenteeksi, joka vastaavansisältöisestä ilmauksesta olisi kohdekielen jäsentimellä tullut tulokseksi. Kyse on lähinnä puurakenteiden muuntamisesta ja vastinilmausten ja sanojen löytämisestä, mitä tehtävää jäsennyksen tuottama rakenne helpottaa.

Paljon käytetty lienee sellainen menettely, jossa lähtökieltä muunnetaan askelittain, kunnes tulos joidenkin askelien jälkeen katsotaan valmiiksi käännökseksi. Menettelyä voidaan kutsua suoraksi kääntämiseksi (engl. direct

translation) siinä mielessä, että vain vähän kääntimen osia voitaisiin hyödyntää missään muussa kieliparissa. Suora kääntäminen esim. japanista englanniksi saatettaisiin tehdä seuraavanlaisten askelien kautta:

1. morfologinen jäsennys (analyysi)
2. sisältösanojen vaihtaminen englanninkielisiksi
3. prepositiorakenteiden sovittaminen
4. lauseiden kieliopillisen sanajärjestyksen korjaaminen kohdekielen mukaiseksi
5. sekalaisia korjailuja
6. taivutusmuotojen generointi

Jos käännettävien tekstien aihepiiri on rajallinen ja mallinnettavissa, voidaan päästä hyvinkin korkealaatuiseen automaattiseen konekäännökseen. Tämä edellyttää sitä, että aihealueen käsitteistä on eksplisiittinen kuvaus, jonka mukaiseen esitysmuotoon lähtökielen jäsennys palauttaa käännettävän lauseen. Vastaavasti generointi tuottaa tällaisesta esitysmuodosta kohdekielelle ilmauksen, joka vastaa tuota abstraktia merkitystä. Aarne Ranta on kehittänyt tällaisen formalismin, jota kutsutaan nimellä *Grammatical Framework* eli GF, jolle on jäsentimiä ja generointisääntöjä usealle kymmenelle kielelle (Ranta et al, 2010). GF-formalismin ja ohjelmien avulla on tehty mm. viimeaikaisessa EU:n komission rahoittamassa MOLTO-hankkeessa.[4] Hintana korkealaatuisesta käännöksestä on

[4] Ks. `http://www.molto-project.eu/`

vaiva, joka tarvitaan aihealueen sääntöjen laatimiseen. MOLTO-hankkeessa pyritään kehittämään menetelmiä, jotka nopeuttavat ja helpottavat tällaisten kielioppien koostamista mm. olemassa olevista käsitehierarkioista ym.

7.5.2 Tilastollinen konekäännös

Sääntöpohjaisen automaattisen kielenkäännön lisäksi on kehitetty *tilastollisen kääntämisen* menetelmiä. Ne perustuvat siihen, että käytettävissä on suuria määriä lähtö- ja kohdekielen aineistoja ja lisäksi *kaksikielistä aineistoa*, joissa sama teksti alkukielellä ja käännettynä jollekin toiselle kielelle. Jos aineistoa on kylliksi ja käytettävissä ehkä vielä *kaksikielinen sanakirja* kielten välille, voidaan käännöksiä muodostaa aineistosta laskettavien tilastojen välille.

Tilastolliset kääntämisen menetelmät ovat varsin houkuttelevia silloin, kun käytettävissä on runsaasti materiaalia. Ajatuksena niissä on, että käännettävien yksiköiden tunnistaminen ja niiden käännösvastineiden identifioiminen tehdään *koneoppimisen* menetelmin tai *tilastollisilla* kriteereillä. Näin katsotaan voitavan säästää kallista ja aikaa vievää ihmistyötä sääntöjen kirjoittamisessa. Tilastollisessa kääntämisessä kohdekielen puolella käytetään ns. *kielimallia*, joka tilastollisin perustein arvioi syntyvän käännöksen todennäköisyyttä eli *kielenmukaisuutta*. Yksittäisten saneiden käännösvaihtoehtojen yhdistelmistä jotkut sopivat paremmin kohdekielen ilmaukseksi kuin toiset. Toisaalta tietyt käännösvastineet ovat lähtökielen saneille tavallisempia kuin toiset. Käännösjärjestelmän tehtävänä on yhdistellä tästä kokonaisuus, joka olisi toisaalta luon-

teva kohdekielessä ja toisaalta todennäköinen yksittäisten sanojen käännöksinä.

Tilastollisessa kääntämisessä ajatellaan yleensä kahta optimoitavaa asiaa: (1) käännöksen pitäisi vastata sanakirjan tai tilastollisesti pääteltyjen käännösvastaavuuksien mukaan *mahdollisimman uskollisesti alkutekstiä* ja (2) tuloksen tulisi olla *mahdollisimman luontevaa kohdekieltä*. Sanojen vastaavuuksia saadaan esimerkiksi sanakirjasta. Kaksikielisestä aineistosta saadaan hyödyllisiä tilastoja näiden vastaavuuksien tarkentamiseksi. Ensin *kohdistetaan* (engl. align) taustalla olevat suuret aineistot virke virkkeeltä käyttäen apuna kappalejakoa ja sanakirjaa. Kohdistetusta aineistosta saadaan parempia arvauksia käännettävässä virkkeessä olevan sanan alamerkityksille. Näin siis voidaan hahmottaa ensimmäistä askelta.

Suuresta aineistosta voidaan laskea myös todennäköisyyksiä sille, miten kohdekielen sanojen sopii *luontevasti seurata toisiaan*. Usein lähtökielen sanan oikea alamerkitys voi selvitä kohdekielenkin puolelta.

Konekääntämisen metodit käsittelevät sitä, millaisia itse asiassa varsin yksinkertaistettuja malleja kielestä pystytään tekemään. Usein tällainen kielimalli perustuu *sanaparien* esiintymisen todennäköisyyksiin, tai ehkä kolmen sanan yhdistelmiin. Kohdekielen, tuloskielen ja käännösvastaavuuksien todennäköisyyksien mallintamisella voidaan kuitenkin päästä käyttökelpoisiin tuloksiin. Tosin siinäkin onnistutaan paremmin, kun kielessä on vain vähän morfologiaa, kuten englannissa.

7.6 Tulevaisuudennäkymiä

Kielen kääntämisen tarve tulee varmaankin kasvamaan entisestään. On houkuttelevaa ajatella, missä kaikkialla voitaisiin hyödyntää nykyistä korkealaatuisempia kääntimiä. Usein vedotaan sellaisiin tulevaisuudenkuviin, joissa puhelimella soittaja voisi puhua vaikka suomea japanilaiselle kollegalleen, joka kuulisi puheen japaniksi, vastaisi japaniksi, joka taas puheentunnistuksen, kääntämisen ja puhesynteesin kautta tulisi selvänä suomena soittajan korvaan. Jos tällaista olisi tarjolla, ihmiset mieluusti varmaan hyödyntäisivät sellaisia sovelluksia. Ne ovat kuitenkin toteutettaviksi niitä kaikkein vaikeimpia, minkä vuoksi emme välttämättä ehdi niitä elinaikanamme nähdä muuta kuin ehkä rajatuissa sovelluksissa.

Liite A

Oheismateriaalia

Kieliteknologian johdantokurssilla materiaali on käyty lävitse kuuden viikon aikana, suunnilleen yksi luku kerrallaan.

Ensimmäinen viikko: johdanto ja kirjoittajan apuvälineet

1. Lue tämän oppikirjan 1. luku eli johdanto.

2. Taustatiedoksi voit lukaista Jurafsky & Martin (2008) -teoksen johdannosta sivut 9–16, joka käsittelee kieliteknologian historiaa.

3. Lue *Oxford Handbook of Computational Linguistics* (Mitkov, 2003) -kirjan luku 2. Teksti on tiivistä, eikä kaikkea oleteta ymmärrettäväksi, sillä asiasta on myöhemmissä opinnoissa erityisiä kursseja. Teksti on tässä taustatiedoksi.

4. Lue tämän oppikirjan 2. luku, joka käsittelee kirjoittajan apuvälineitä.

Toinen viikko: tiedonhaku

1. Lue tämän oppikirjan 3. luku eli "tiedonhaku".

2. Lue *Oxford Handbook of Computational Linguistics* (Mitkov, 2003) -kirjan luku 29. Teksti on tiivistä, eikä aivan kaikkea oleteta ymmärrettäväksi, sillä teemasta on myöhemmissä opinnoissa erityisiä kursseja. Teksti on tarkoitettu tässä vaiheessa yleiskatsaukseksi, joka täydentää tämän oppikirjan tietoja.

3. Jos et ennestään ole perehtynyt tiedonhakuun (esim. opiskelemalla informaationtutkimusta Tampereella), tutustu kirjaan *Information Retrieval* (Van Rijsbergen, 1979). Lue ainakin kursorisesti sen kirjan luvut: 1: *Introduction* ja 5: *Search Strategies*. Kirjasta lienee vapaasti verkossa luettava versio.

Kolmas viikko: puheteknologia

1. Lue tämän oppikirjan 4. luku, joka käsittelee puheteknologiaa.

2. Lue *Oxford Handbook of Computational Linguistics* -kirjan luku 16 *Speech recognition* ja luku 17 *Text-to-speech synthesis*. Teksti on tiivistä, eikä aivan kaikkea oleteta ymmärrettäväksi, sillä asiasta on kaksikin myöhempää kurssia. Teksti on tässä yleiskatsaukseksi, joka täydentää kurssimonistetta.

3. Katsele paria kuvaa kirjasta *Fundamentals of speech recognition* (Rabiner & Juang, 1993) kirjasta: kuva 2.6 (jossa mekaaninen kaavakuva puhetta tuottavista elimistä) sivulla 17, kuva 2.13 ääntöväylän asennoista sivulla 25 ja kuvat 2.15 ja 2.16 (vokaalien spektrogrammeja, ja vokaalien hajontaa ja jakautumista formanttien perusteella) sivulla 27. Ellet saa kirjaa käsiisi, tutki englanninkielistä Wikipediaa, jossa voi olla vastaavia kuvia selityksineen.

4. Tarkastele kirjasta (Jurafsky & Martin, 2008) neljännessä luvussa mainittuja kuvia, erityisesti niitä, jotka havainnollistavat äänen aaltomuotoa, spektrogrammia ja äänen hetkellistä spektriä. Ellet saa kirjaa käsiisi, etsi Wikipediasta vastaavia.

5. Tutustu johonkin verkossa vapaasti saatavaan puhesynteesin demoon, esimerkiksi Bitlips Oy:n demoihin (`http://www.bitlips.fi/`), joissa on erilaisia selaimen kautta käytettäviä suomen kielen puhesyntetisaattoreita.

Neljäs viikko: vuorovaikutus ihmisen ja koneen välillä

1. Lue tämän oppikirjan 5. luku, joka käsittelee luonnollisella kielellä tapahtuvaa vuorovaikutusta ihmisen ja koneen välillä.

2. Lue artikkelista *Spoken Dialogue Technology: Enabling the Conversational User Interface* (McTear, 2002): sivut 90-104, tarkemmin luettavaksi ja pääkohdat ym-

märrettäväksi, 113-114, kohdan 4.3 alku, erityisesti kuva 7. Artikkelin lopulla s. 162-163 on valikoima alan tutkimushankkeista ja saatavilla olevista työkaluista, jotka on myös syytä selata lävitse. Muuhun osaan artikkelia ei ole tarvis kajota tällä kurssilla (eikä tulostaa paperille). Voit lukea vastaavan materiaalin myös McTearin myöhemmästä kirjasta (McTear, 2004).

3. Lue *Oxford Handbook of Computational Linguistics* (Mitkov, 2003) luku 31 *Question answering*. Teksti on tiivistä, eikä aivan kaikkea oleteta ymmärrettäväksi. Teksti on tässä yleiskatsaukseksi, joka täydentää kurssimonistetta.

4. Vilkaise seuraavia Turingin testiin liittyviä paikkoja (jos löydät ne, muuten voit sivuuttaa ne):

   - Turingin testin kotisivu `http://www.fil.ion.ucl.ac.uk/~asaygin/tt/ttest.html`. Siellä on kiinnostuneelle lukijalle paljon hyviä linkkejä ja viitteitä sekä Turingin testiä, sen tulkinnasta että esim. demo-ohjelmista.
   - Historiallisestikin kiinnostava on Turingin testiä koskeva alkuperäisen artikkelin kopio `http://cogprints.org/499/1/turing.html`, kun otetaan huomioon artikkelin kirjoittamisen ajankohta ennen kuin varsinaisia tietokoneita oli juurikaan käytettävissä.

- Vuosittain pidettävän kilpailun ja sen ns. Loebnerin palkinnon kotisivu `http://www.loebner.net/Prizef/loebner-prize.html`, jolla on myös linkkejä ja muuta hyödyllistä tietoa katsokaa erityisesti tämän vuoden voittajaohjelmien keskusteluiden transkriptioita (ovat aika vakuuttavia).
- Klassinen artikkeli psykiatriohjelmasta: *ELIZA–A Computer Program For the Study of Natural Language Communication Between Man and Machine* (Weizenbaum, 1966) kuvailee ohjelman.
- Wikipediassa on tietoa klassillisista keskusteluohjelmista ELIZA, joka jäljittelee Carl Rogersin tyylistä terapeuttia, ja PARRY, joka on paranoiaa jäljittelevä ohjelma, ks. *Artificial Paranoia* (Colby et al, 1971).
- Vähän toisenlainen keskustelija, jonka kanssa voi jutella mistä vain: ALICEBOT eli Loebnerin palkinnon voittaja vuodelta 2004. `http://alice.pandorabots.com/`
- Seuraavilla sivulla on runsaasti linkkejä erilaisiin keskusteluohjelmiin: (a) The Simon Laven Page, (b) Wikipedian *Chatterbot* -artikkeli ja (c) ja erikseen on mainittava *Albert One*, joka voitti Loebnerin palkinnon 1998 ja 1999.

Viides viikko: kieliteknologia ja kielen oppiminen

1. Lue kurssimonisteen 6. luku joka käsittelee kielenoppimisen ja -opettamisen kieliteknologiaa.

2. Lue opetusmonisteen *Datorstödd språkinlärning och språkteknologi* (Borin, 2005) kolmannesta luvusta *AI i CALL program: ICALL* kohta 3.2 *Datamaskinell språkanalys (parsning) och grammatikkontroll i CALL*, kohta 3.3 *Informationsädling i ICALL*, kohta 3.4 *Inlärningskorpusar*, kohta 3.5. *Inlärarmolellering i ICALL*, s. 40–53. Jos et saa tätä käsiisi tai et ymmärrä yhtään ruotsia, jätä väliin.

3. Lue *Oxford Handbook of Computational Linguistics* (Mitkov, 2003) luku 37 *Natural language processing in computer-assisted language learning* ja siitä erityisesti s. 680-690 eli kohta 37.6 *Contributing NLP Technologies*. Muu osa sisältää aihepiirin motivointia ja historiaa. Luettava tekstin osa on tiivistä, eikä aivan kaikkea oleteta ymmärrettäväksi. Teksti on tarkoitettu yleiskatsaukseksi, joka täydentää kurssimonistetta.

4. Edellisen sijasta tai täydennyksenä voit lukea Wikipedian Computer assisted language learning -artikkelin, jossa on käsitelty kieliteknologiankin osuutta CALLissa sekä annettu koko joukko viitteitä ja linkkejä ynnä hyvä yleiskuva aihepiiristä.

5. Varsinkin, jos olet kiinnostunut asiasta, voit vilkaista seuraavia verkkosivuja, joissa on mielenkiintoisia

aihepiiriin liittyviä tietoja ja linkkejä:

- Claire Bradin Siskinin (luennoitsija Pittsburghin yliopistosta) keräämää CALL linkkisivua
  http://edvista.com/claire/call.html
- Warschauer, M. (1996). Computer-assisted language learning: An introduction. In S. Fotos (Ed.), Multimedia language teaching (pp. 3-20). Tokyo: Logos International.
- LOCOLEX- (Segond et al, 1995) ja GLOSSER-RuG (Nerbonne & Smit, 1996) (verkossa).
- BootCaT (Baroni & Bernardini, 2004): Bootstrapping corpora and terms from the web. Proceedings of LREC 2004. Ks. myös verkkosivua.[1]
- Vokaalipeli (Paganus et al, 2006, pp. 696-703).
- Pohjoissaamen opiskelua varten laadittu sivusto, jolla on useita erilaisia harjoituksia ja pelejä, http://oahpa.no/davvi/. Muutamat niistä perustuvat pohjoissaamen morfologiseen jäsentimeen sekä syntaktiseen jäsentimeen, joka yksiselitteistää morfologisesti moniselitteisiä sananmuotoja.
- Eckard Bickin kehittämä VISL (Visual Interactive Syntax Learner) -sivusto,
  http://beta.visl.sdu.dk/
  jolla erilaisia tietokonepelejä, jotka on sovitettu kieliopillisten käsitteiden oppimista varten.

[1] http://bootcat.sslmit.unibo.it/

Niiden avulla opetellaan mm. sanaluokkia ja lauseenjäseniä ampumalla niitä siihen tapaan, kuin videopeleissä ammutaan tunkeutujien avaruusaluksia tms.

Kuudes viikko: kielen kääntämisen apuvälineet

1. Lue kurssimonisteen 7. luku, joka käsittelee Kielen kääntämisen apuvälineitä ja automaattista kielen kääntämistä.

2. Lukekaa kirjasta (Arnold et al, 1993) luku 1, Introduction and Overview, luku 2 Machine Translation in Practice, ks. `http://www.essex.ac.uk/linguistics/external/clmt/MTbook/PostScript/`

3. Selailkaa historiallisesti merkittävää ALPAC -komitean raporttia, jonka sanotaan katkaisseen konekäänTämisen runsaskätisen rahoituksen 1960-luvun puolivälissä. Selaile mm. sivuilla 19-24 oleva *The Present State of Machine Translation*, sivulla 34 olevaa *Recommendations*, sekä sivuilla 29-31 olevaa lukua *Automatic Language Processing and Computational Linguistics*, joka ei ole ollenkaan niin vanhanaikainen, kuin kirjoitusvuodesta 1966 saattaisi epäillä.

4. Muuta taustatietoa:

   - Kokonainen kirja johdatuksena konekääntämiseen (Hutchins & Somers, 1992).

- Kielikone Oy:n MOT-Sanakirjasto, joka on käytettävissä ainakin Helsingin yliopiston verkosta käsin (esim. Alman kautta).
- Mickel Grönroosin kirjoitus Från översättningsminne till översättningsintelligens jossa kerrotaan kieliteknologialla varustetusta käännösmuistista ja perustellaan sellaisen hyödyllisyyttä, http://www.kotus.fi/index.phtml?l=sv&s=591
- Ruotsin Skoldatanätet -hankkeen LEXIN -sanakirjapalvelu, jossa lukuisia vapaasti verkossa käytettäviä sanakirjoja, mm. ruotsi-suomi-ruotsi, http://lexin2.nada.kth.se/lexin/.

Kirjallisuutta

ALPAC, 1966. *Language and Machines: Computers in Translation and Linguistics*. Publication 1416. The National Academy of sciences, Automatic Language Processing Advisory Committee. Saatavissa: http://www.nap.edu/openbook.php?record_id=9547.

D.J. Arnold, Lorna Balkan, Siety Meijer, R.Lee Humphreys, Louisa Sadler, 1993. *Machine Translation: an Introductory Guide*. Blackwells-NCC, London. http://www.essex.ac.uk/linguistics/external/clmt/MTbook/PostScript/.

M. Baroni, S. Bernardini, 2004. "Bootcat: Bootstrapping corpora and terms from the web". Teoksessa *Proceedings of LREC 2004, Lisbon*, ss. 1313–1316. ELDA. http://www.cs.utah.edu/nlp/readinglist/BaroniB04.pdf.

Kenneth R. Beesley, Lauri Karttunen, 2003. *Finite State Morphology*. Studies in Computational Linguistics, 3. University of Chicago Press. Kirjaa täydentäviä tietoja,

katso: www.stanford.edu/~laurik/fsmbook/home.html.

Daniel G. Bobrow, Ronald M. Kaplan, Martin Kay, Donald A. Norman, Henry S. Thompson, Terry Winograd, 1977. “Gus, a frame-driven dialog system”. *Artificial Intelligence*, 8 (2):155–173.

Lars Borin, 2005. “Datorstödd språkinlärning och språkteknologi. kurskompendium. (fjärde upplagan.)”. kursmaterial, Institutionen för svenska språket, Göteborgs universitet.

K.M. Colby, S. Weber, F.D. Hilf, 1971. “Artificial paranoia”. *Artificial Intelligence*, 2:1–25.

Jonathan Harrington, Steve Cassidy, 1997. *Techniques in Speech Acoustics*. Kluwer Academic Publishers.

James Hillenbrand, Laura A. Getty, Michael J. Clark, Kimberlee Wheeler, 1995. “Acoustic characteristics of american english vowels”. *Journal of the Acoustical Society of America*, 97 (5, Pt. 1):3099–3109.

W. John Hutchins, Harold L. Somers, 1992. *An introduction to machine translation*. Academic Press.

Daniel Jurafsky, James H. Martin, 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Series in Artificial Intelligence. Pearson Prentice Hall.

Fred Karlsson, Kimmo Koskenniemi, 1990. *BETA-ohjelma kielentutkijan apuvälineenä*. Yliopistopaino, Helsinki.

Kimmo Koskenniemi, Krister Lindén, Lauri Carlson, Martti Vainio, Antti Arppe, Mietta Lennes, Hanna Westerlund, Mirka Hyvärinen, Imre Bartis, Pirkko Nuolijärvi, Aino Piehl, 2012. *Suomen kieli digitaalisella aikakaudella – The Finnish Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Saatavissa: http://www.meta-net.eu/whitepapers.

Timo Lahtinen, 2000. *Automatic indexing: an approach using an index term corpus and combining linguistic and statistical methods*. Numero 34 teoksessa Publications. University of Helsinki, Faculty of Arts, Department of General Linguistics. Saatavissa: http://ethesis.helsinki.fi/julkaisut/hum/yleis/vk/lahtinen/automati.pdf.

M. Paul Lewis (toimittaja), 2009. *Ethnologue: Languages of the World*. SIL International, Dallas, Tex., 16. painos. Selattavissa: http://www.ethnologue.com/.

Michael McTear, 2002. "Spoken dialogue technology: enabling the conversational user interface". *ACM Comput. Surv.*, 34 (1):90–169. Http://doi.acm.org/10.1145/505282.505285.

Michael McTear, 2004. *Spoken Dialogue Technology: Toward the Conversational User Interface*. Springer.

Ruslan Mitkov, 2003. *The Oxford Handbook of Computational Linguistics*. Oxford Handbooks in Linguistics. Oxford University Press.

John Nerbonne, Petra Smit, 1996. "Glosser-rug: in support of reading". Teoksessa *Proceedings of the 16th conference on Computational linguistics*, osa 2, ss. 830–835. COLING, Copenhagen. http://www.aclweb.org/anthology/C96-2140.

Annu Paganus, Vesa-Petteri Mikkonen, Tomi Mäntylä, Sami Nuuttila, Jouni Isoaho, Olli Aaltonen, Tapio Salakoski, 2006. "The vowel game: Contrinuous real-time visualization for pronunciation learning with vowel charts". Teoksessa *Advances in Natural Language Processing, 5th International Conference, FinTAL 2006 Turku, Finland, August 23-25, 2006 Proceedings*, osa 4139 sarjasta *Lecture Notes in Computer Science*, ss. 696–703. Springer Berlin / Heidelberg.

Lawrence Rabiner, Biing-Hwang Juang, 1993. *Fundamentals of speech recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Aarne Ranta, Krasimir Angelov, Thomas Hallgren, 2010. "Tools for multilingual grammar-based translation on the web". Teoksessa *Proceedings of the ACL 2010 System Demonstrations*, ss. 66–71. Association for Computational Linguistics, Uppsala, Sweden.

Saatavissa: http://www.aclweb.org/anthology/P10-4012.

Frederique Segond, Daniel Bauer, Annie Zaenen, 1995. “Locolex: Translation rolls off your tongue”. Teoksessa *ACH-ALLC 95 Proceedings*. AC-ALLC.

C.J. Van Rijsbergen, 1979. *Information retrieval*. Butterworths. Katso: http://www.dcs.gla.ac.uk/Keith/Preface.html.

Reetta Vuokko, 2009. *Applying Language Technology to Computer-Assisted Language Learning: Producing Microcorpora for Early Learning Stages of Japanese*. pro gradu, University of Helsinki, Language technology.

Joseph Weizenbaum, 1966. “Eliza: a computer program for the study of natural language communication between man and machine”. *Commun. ACM*, 9 (1):36–45. http://doi.acm.org/10.1145/365153.365168.

Hakemisto

Nykykielten laitoksen oppimateriaalia

Nykykielten laitoksen oppimateriaalia on Helsingin yliopiston nykykielten laitoksen (http://www.helsinki.fi/nykykielet) julkaisema verkkojulkaisusarja, joka sisältää laitoksen opettajien laatimaa niin opiskelijoille kuin laajemmallekin yleisölle suunnattua korkealaatuista oppimateriaalia kielen- ja kirjallisuudentutkimuksen eri osa-alueilta.

ISSN 2323-8828 (Verkkojulkaisu)

hdl.handle.net/10138/38336

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI