Aalto-yliopisto
Perustieteiden korkeakoulu
Informaatioverkostojen koulutusohjelma

Mallinnusyksikön valinta tilastollisissa kielimalleissa

Kandidaatintyö

28. marraskuuta 2011

Noora Routasuo

Aalto-yliopisto
Perustieteiden korkeakoulu
Informaatioverkostojen koulutusohjelma

KANDIDAATINTYÖN
TIIVISTELMÄ

Tekijä:	Noora Routasuo
Työn nimi:	Mallinnusyksikön valinta tilastollisissa kielimalleissa
Päiväys:	28. marraskuuta 2011
Sivumäärä:	24
Pääaine:	Mediatekniikka
Koodi:	T3004
Vastuuopettaja:	Ma professori Stina Immonen
Työn ohjaaja(t):	DI Sami Virpioja (Tietojenkäsittelytieteen laitos)

Tilastollisessa kielen mallinnuksessa käytetään mallinnusyksikkönä useimmiten sanaa, jolloin mallin todennäköisyydet lasketaan sanoista koostuvien sekvenssien perusteella. Toisaalta yksikkönä voidaan käyttää myös yhtä sanaa lyhempää tai pidempää yksikköä. Tämän kandidaatintyön tavoitteena on kartoittaa seikkoja, jotka vaikuttavat mallinnusyksikön valintaan. Työssä käsitellään n-grammipohjaisia kielimalleja ja sovelluksista puheentunnistusta ja konekäännöstä. Tutkimus pohjautuu julkaistuihin mallinnustuloksiin sekä kielitieteen teoriaan.

Mahdollisia yksiköitä ovat tavu, kieliopillinen ja tilastollinen morfi, sana sekä fraasi. Niiden piirteitä voidaan myös yhdistellä hierarkisiksi tai faktorimalleiksi. Hyvä mallinnusyksikkö on sellainen, josta voidaan muodostaa pienelläkin aineistolla kattava, mutta toisaalta isollakin aineistolla käytännöllisen kokoinen malli. Yleisesti lyhyempi yksikkö mahdollistaa pienemmän aineiston käytön, mutta vaatii pidempiä n-grammeja ja voi siten kasvattaa mallin kokoa.

Yksikön valintaan vaikuttaa kolme päätekijää: mallinnettava kieli, sovellus sekä käytössä olevat resurssit. Parhaat mallinnustulokset on saatu eri kielissä ja eri sovelluksissa eri yksikköä käyttäen, mutta tilastollisin, aineistosta opituin yksiköin on saatu hyviä tuloksia myös kielestä tai sovelluksesta riippumatta.

Tulosten perusteella morfologisesti monimutkaisemmat kielet vaativat sanaa pienemmän pienemmän mallinnusyksikön. Sovelluksista kielenkäännöksessä fraasi on lupaava yksikkö, kun taas etenkin laajan sanaston puheentunnistuksessa lyhyemmällä yksiköllä on saatu parhaat tulokset. Sovelluskohtaiset tulokset riippuvat kuitenkin tarkemmasta tehtävänmäärittelystä. Resurssit kuten käytettävissä oleva aineisto ohjaavat yksikön valintaa siten, että lyhyemmällä yksiköllä voidaan hyödyntää pienempää aineistoa.

Avainsanat:	kielimalli, n-grammit, luonnollisen kielen mallinnus, puheentunnistus, konekäännös, mallinnusyksikkö
Kieli:	Suomi

Aalto University
School of Science
Degree Program of Information Networks

ABSTRACT OF
BACHELOR'S THESIS

Author:	Noora Routasuo
Title of thesis:	Selection of the modelling unit in statistical language models
Date:	November 28, 2011
Pages:	24
Major:	Mediatekniikka
Code:	T3004
Supervisor:	Professor (pro tem) Stina Immonen
Instructor:	Sami Virpioja, M.Sc. (Tech.) (Department of Information and Computer Science)

Statistical language modelling most commonly uses words as the basic units of modelling. Probabilities in the model are calculated based on sequences of words. However, units shorter or longer than word can also been used. The aim of this work is to investigate the factors affecting the choice of the modelling unit. The work focuses on n-gram models and particularly the applications of machine translation and automatic speech recognition, and is based on published modelling results and linguistic theory.

Possible modelling units include syllables, linguistic and statistical morphs, words, and phrases. Their features can also be combined in hierarchical or factored models. A good modelling unit enables the construction of a model with a good coverage even with a small amount of training data while still keeping the size of the model practical with a larger amount. In general, shorter units require less data, but also longer n-grams which may make the resulting model bigger.

The three main factors contributing to the choice of the modelling unit are the language to be modelled, the application, and the available resources. The best results have been achieved using different units in different languages and applications, but statistical methods have proven to be portable across languages and applications under certain conditions.

Based on the reviewed publications, morphologically complex languages generally require a unit shorter than word. Phrase is a promising unit in machine translation, whereas especially in large vocabulary speech recognition the best unit is often short. However, the best unit for each application seems to depend on a more specific context. As for available resources, the most important factor is size of the corpus.

Keywords:	language model, n-gram models, natural language processing, speech recognition, machine translation, modelling unit
Language:	Finnish

Sisältö

1 Johdanto

Kaikissa tilanteissa, joissa halutaan koneen tuottavan, ymmärtävän tai muuten käsittelevän automaattisesti luonnollista kieltä vaaditaan kielen mallinnusta. Sovelluksia ovat muun muassa puheentunnistus (*speech recognition*), konekäännös (*machine translation*) ja tiedonhaku (*information retrieval*). Tavallisin lähestymistapa ovat tilastolliset kielimallit, jotka mallintavat kieltä todennäköisyyksien avulla. Tilastollisessa mallinnuksessa kullekin kielen yksikölle, kuten sanalle, lasketaan todennäköisyys, kun sen esiintymisestä ja kontekstista tiedetään jotakin.

Tämä kandidaatintyö käsittelee mallinnusyksikön valintaa kielen tilastollisessa mallinnuksessa. Tyypillisimmin yksikkönä käytetään sanaa, mutta joissakin tilanteissa sanaa lyhyempi tai jopa pidempi yksikkö voi olla paremmin perusteltu vaihtoehto. Esimerkiksi mallinnettaessa vahvasti taipuvia kieliä, kuten suomea, sanojen käsitteleminen sellaisenaan johtaa mahdollisen sanaston kasvamiseen epäkäytännöllisen suureksi. Toisaalta konekäännöksessä sanoja pidemmät fraasit usein ratkaisevat käännöksen sujuvuuden.

Työn tavoitteena on selvittää, miten yksikön valinta vaikuttaa kielimallin toimivuuteen erilaisissa tilanteissa, ja siten luoda käsitys niistä seikoista, jotka vaikuttavat yksikön valitaan. Kysymystä tutkitaan sekä kohdekielen että mallinnuksen käyttötarkoituksen kannalta. Tässä työssä tarkastelun kohteena ovat erityisesti suomi ja sen sukulaiskielet sekä thai, arabia ja turkki. Sovelluksista on keskitytty pääasiassa puheentunnistukseen ja konekäännökseen, ja mallinnustekniikoista n-grammipohjaisiin malleihin.

Tutkimuksen perustana ovat mallinnuksen olemassa olevat ratkaisut eri kielissä ja eri sovellusalueilla. Näitä ratkaisuja verrataan keskenään yksikön valinnan näkökulmasta sikäli kun se on mahdollista. Lisäksi tarkastellaan kielien ja sovellusten erityispiirteitä yksikön valinnan näkökulmasta hyödyntäen kielitieteen teoriaa.

Yksikön valintaa on lähestytty tutkimuksessa pääasiassa vertailemalla sanoja fraaseihin tai morfeihin mallinnusyksiköinä. Paras yksikkö on ainakin jossakin määrin kieli- ja sovellusriippuvainen, mutta myös näistä riippumattomilla ratkaisuilla ollaan saatu hyviä tuloksia. Kielen kannalta tärkein yksikönvalintaan vaikutavat tekijä on sen morfologisen rakenteen monimutkaisuus. Eri sovelluksissa paras yksikkö riippuu monista eri tekijöistä. Käytössä olevat resurssit sen sijaan vaikuttavat valintaan siten, että lyhyemmillä yksiköillä voidaan saada hyviä tuloksia pienemmästäkin aineistosta.

Työn rakenne on seuraava: luvussa 2 esitellään kieliteoreettinen tausta, kielen mallinnuksen perustekniikat sekä mallien vertailukriteerit. Yksikön rooli kielimalleissa ja erilaisia yksiköitä käsitellään luvussa 3. Yksikön valintaan vaikuttavia seikkoja eritellään yksitellen luvussa 4. Luku 5 käsittelee yksikönvalintaa kokonaiskysymyksenä ja yhteenveto tehdään luvussa 6.

2 Tausta

Tässä luvussa käydään läpi kielitieteellinen tausta, kielen tilastollisen mallinnuksen perustekniikat sekä kielimallien vertailukriteereitä.

2.1 Kielitieteen peruskäsitteitä

Vaikka tilastollinen kielen mallinnus hyödyntääkin usein melko vähän tietoa kielen erityisrakenteesta, on joihinkin malleihin pyritty sisällyttämään myös kielitieteellistä tietoa kielen säännönmukaisuuksista. Erityisesti mallinnuksen yksikön määritteleminen ja kielten vertailu keskenään vaatii kielitieteellisen pohjan, joka annetaan tiivistetysti tässä osiossa.

2.1.1 Luonnollisen kielen osat

Ihmiskieli on säännönmukainen järjestelmä, josta voidaan erottaa joitakin kaikille kielille yhteisiä osia. Kaikki kielet koostuvat sanoista ja lauseista, mutta sanan määritelmä ei aina ole aivan yksiselitteinen. Aakkosin kirjoitetun kielen mallinnuksessa sanana käsitellään yleensä yksinkertaisesti välein erottuvia osia, jolloin esimerkiksi yhdyssanat käsitellään yhtenä sanana. Lekseemi (*lexeme*) on sanan abstraktio, kaikkien saman perusmuodon, lemman, ilmentymien eli pintamuotojen joukko.

Sanat koostuvat morfeemeista, jotka ovat kielen pienimpiä merkitysyksiköitä. Morfeemien ilmentymiä, kirjoitusasultaan keskenään erilaisia, mutta merkitykseltään samoja palasia kutsutaan morfeiksi. Esimerkiksi englannin monikko `-s` on morfeemi, joka ilmenee sanasta riippuen muun muassa morfeina `-s` ja `-es`. Sanojen jako morfeihin ei ole aina yksiselitteinen, mutta yleensä melko intuitiivinen (Karlsson, 2008). Esimerkiksi sana `koirallekin` voidaan jakaa itsenäisen merkityksen omaaviin osiin `koira`, `-lle` ja `-kin`. Sana voidaan myös jakaa vartaloon (*stem*) ja erilaisiin liitteisiin eli affikseihin, joita ovat vartalon eteen tulevat prefiksit, päätteet eli suffiksit ja keskelle vartaloa lisättävät infiksit.

Sanat voivat muodostua morfeemeista ja lekseemeistä lukuisin eri tavoin. Luonnollisen kielen käsittelyn kannalta keskeisimpiä näistä ovat taivutus, johdokset ja yhdyssanat. Taivutus on lekseemien muovaamista lisäämällä niihin esimerkiksi persoonaan, lukumäärään tai aikaan liittyvää informaatiota. Tämä tapahtuu useimmiten taivutuspäätteillä. Johdokset ovat säännöllisellä tavalla olemassa olevista lekseemeistä muodostettuja uusia, itsenäisiä lekseemejä. Esimerkiksi sanat `kalastaa` ja `kalamainen` ovat sanan `kala` johdoksia. Yhdyssanat ovat itsenäisistä sanoista muodostettuja uusia sanoja, joilla on niiden osiin perustuva merkitys. Kielestä riippuen yhdyssanat kirjoitetaan yhteen tai erikseen.

Puhutun kielen pienin merkityksiä erotteleva yksikkö on foneemi, ja puhetta voidaan kuvata foneemien sarjana. Foneemi on abstraktio, sillä lausumiseen vaikuttavat luke-

mattomat seikat (puhuja, ympäröivät äänteet ja niin edelleen). Kun vaihtelu ei aiheuta epäselvyyttä merkityksen suhteen, kyse on saman foneemin eri ilmentymistä eli fooneista. Monissakaan kielissä äänteet eivät vastaa läheskään yhtä suoraviivaisesti kirjoitettua tekstiä kuin suomessa.

Tavu ei ole tarkasti määritelty käsite, mutta sen ajatellaan yleisimmin sisältävän vokaalin ja joitakin siihen läheisesti liittyviä konsonantteja (Jurafsky ja Martin, 2008). Se on siis myös äänteellinen yksikkö.

2.1.2 Kielten luokittelu

Morfologinen typologia on 1800-luvulla alunperin kehitelty kielten luokittelu (Karlsson, 2008). Sen luokat eivät ole tarkasti rajattuja, vaan sitä pidetään pikemminkin janana, jossa yksittäiset kielet sijoittuvat jonnekin eri kategorioiden välille.

Kieliä, joissa yksi sana sisältää on tyypillisesti vain yhden morfeemin, kutsutaan analyyttisiksi (*analytic*) tai isoloiviksi. Niissä kieliopillisia suhteista ilmaistaan omilla sanoillaan, sanat eivät juurikaan taivu ja sanajärjestys on siten tärkeä. Tyypillisiä hyvin isoloivia kieliä ovat kiina ja thai.

Kieliä, joissa sana sisältää tyypillisesti useita morfeemeja kutsutaan synteettisiksi (*synthetic*). Nämä jaetaan edelleen kahteen pääluokkaan sen mukaan, erottuvatko morfeemit helposti toisistaan vai sulautuvatko ne yhteen. Agglutinatiivisiksi (*agglutinative*) kutsutaan kieliä, joissa morfeemit erottuvat selkeästi ja yksi morfeemi vastaa karkeasti yhtä merkitystä. Tyypillisiä esimerkkejä ovat uralilaiset kielet kuten suomi ja unkari, sekä altailaiset kielet kuten turkki. Kun morfeemit sekoittuvat keskenään ja aiheuttavat enemmän äänteellistä vaihtelua, puhutaan fuusiokielistä (*fusional language*). Tästä esimerkkejä ovat muun muassa venäjä ja puola.

2.2 Tilastollinen kielen mallinnus

Kielen mallinnuksessa sovelletaan informaatioteoriasta lähtöisin olevaa ajatusta kohinaisesta kanavasta (*noisy channel*). Syötteenä saadaan jokin signaali – esimerkiksi puhetta tai vieraskielinen lause – jonka ajatellaan kulkeneen kohinaisen kanavan läpi. Tehtävänä on selvittää, mikä alkuperäinen, kohinaton signaali oli mallintamalla tätä kohinaista kommunikaatiokanavaa eli kieltä. (Jurafsky ja Martin, 2008)

Tilastollisessa lähestymistavassa kielestä rakennetaan malli ehdottomien sääntöjen sijaan todennäköisyyksien avulla. Pyritään siis määrittelemään, millainen syöte on todennäköisimmin ollut ennen altistumistaan kohinalle, kun lopputulos tiedetään. Todennäköisin

alkuperäinen sana $\hat{w}$, kun kohinainen sana on w ja koko havainto on O, saadaan yhtälöstä:

$$\hat{w} = \arg\max_{w} P(w|O)$$

Todennäköisyyden $P(w|O)$ laskemiseen käytetään Bayesin kaavaa:

$$P(w|O) = \frac{P(O|w)P(w)}{P(O)}$$

Nimittäjä $P(O)$ on maksimointitehtävän kannalta merkityksetön, sillä se on sama kaikille sanoille. Todennäköisyys $P(O|w)$ on sovelluskohtainen, ja sen estimointiin on erilaisia menetelmiä. Puheentunnistuksessa se lasketaan akustisen mallin ja konekäännöksessä käännösmallin avulla. Todennäköisyyttä $P(w)$ kutsutaan prioriksi, ja se saadaan kielimallista. Yksinkertaisimmillaan esitettynä tilastollinen kielimalli on siis todenäköisyysjakauma kaikille mahdollisille mallinnuksen yksikön – yleisimmin sanojen – sekvensseille.

Kielen monimuotoisuudesta seuraa, että minkä tahansa mallin opetukseen tarvitaan suuri määrä aineistoa. Tämä aineisto voi olla merkitsemätöntä tai sisältää erilaista tietoa esimerkiksi sanojen merkityksistä tai sanaluokista. Tällöin puhutaan ohjatuista (*supervised*) ja ohjaamattomista (*unsupervised*) oppimisalgoritmeista. Tyypillisesti aineiston sopivuus on kriittinen tekijä toimivan mallin opetuksessa (Jurafsky ja Martin, 2008). Jos opetusaineiston kieli tai genre poikkeaa suuresti siitä materiaalista, johon mallia lopulta sovelletaan, se harvoin toimii hyvin.

Zipf'in lain mukaan missä tahansa kielessä tai aineistossa on joitakin hyvin yleisiä sanoja ja paljon hyvin harvinaisia sanoja (Jurafsky ja Martin, 2008). Täten kielen kohtalaiseen kattamiseen tarvitaan vain joitakin tavallisimpia sanoja, mutta kattavuuden parantaminen vaiketuu sitä mukaa, mitä lähemmäksi täydellisyyttä pyritään. Tämän vuoksi aineiston harvuus on mallinnuksen tyypillinen ongelma. Hyvät tulokset vaativat aina suuria oikeantyyppisiä aineistoja, joiden kokoaminen, löytäminen ja käsitteleminen ovat kaikki aikaavieviä tehtäviä. Koska mallin opetuksen ulkopuolelle jää aina tuntemattomia sanoja, ne täytyy ottaa erikseen huomioon tavalla tai toisella. Tuntemattomien sekvenssien ongelmaan on erilaisia ratkaisuja, joita ovat muun muassa *smoothing* ja *clustering* -tekniikat (muun muassa Chen ja Goodman, 1999). Tuntemattomien sanojen (*out-of-vocabulary words*) ongelmaa voidaan myös lähestyä käyttämällä mallinnuksessa tietoa kielen rakenteesta – esimerkiksi sanaa pienempiä mallinnusyksiköitä.

Tilastollinen mallinnus on kehittynyt viime vuosina suurempien aineistojen ja paremman laskentatehon myötä. Goodman (2001) kuitenkin huomauttaa, että kehityksen vaikutelma perustuu osin ontuviin vertailumenetelmiin ja käytännön sovelluksien kannalta liian raskaisiin tai monimutkaisiin kokeisiin. Rosenfeld (2000) esittää, että kielimalleihin pitäisi sisällyttää enemmän kielellistä tietoa. Tätä onkin tutkittu viime vuosina, osin uusien tutkittavien kielten vaikutuksen ansiosta (muun muassa Creutz, 2006; Ablimit et al., 2010; Mihajlik et al., 2010).

2.3 N-grammipohjaiset kielimallit

Tavallisimmat tilastollisessa kielen mallinnuksessa käytetyt mallit ovat niin kutsutut n-grammimallit (*n-gram*) ja niiden erilaiset variaatiot. N-grammimallissa yksinkertaistetaan yhden sanan (tai muun yksikön) todennäköisyyden laskemista käyttämällä sen koko historian h_i sijasta vain edeltävää $n-1$ sanaa. Toisin sanoen käytetään ehdollista todennäköisyyttä

$$P(w_i|h_i) = P(w_i|w_{i-n+1}, w_{i-n+2}, ..., w_{i-1})$$

Tämä perustuu intuitiiviseen ajatukseen siitä, että kielen säännönmukaisuudet määräävät millaisia sanoja voi esiintyä peräkkäin. Samalla tehdään kuitenkin oletus, että mallinnettavan sanan todennäköisyys on riippumaton kaikista muista tekstin ominaisuuksista, mikä ei tietenkään pidä paikkaansa. Käytännössä mallien on kuitenkin osoitettu toimivan melko hyvin tästä yksinkertaistuksesta huolimatta (Goodman, 2001).

N-grammin pituus määrää sen ennustuskyvyn, mutta toisaalta tekee mallista myös raskaamman. Kun n kasvaa, saadaan teoriassa täsmällisempiä arvioita, mutta estimointi heikkenee aineiston vähyyden vuoksi (Sarikaya et al., 2008) – vaikka tietty n sanan pituinen sekvenssi olisikin suhteessa yleisempi kuin jokin toinen, sen esiintyminen sellaisenaan on silti epätodennäköisempää kuin lyhyemmän sekvenssin. Rosenfeld (2000) arvioi, että käytännöllisin pituus on kaksi historiasanaa, mutta tietyin oletuksin jopa viiden sanan käyttäminen voi tuottaa malliin parannuksia. Etenkin pienempien yksiköiden kohdalla pidemmän historian käyttö on tarpeellista (Hirsimäki et al., 2009).

Suurilla n ja isoilla aineistoilla mallin koko kasvaa ongelmalliseksi, mutta sitä voidaan rajoittaa erilaisilla tekniikoilla. Karsimisessa (*pruning*) vain yleisimmät tai tulosten kannalta merkityksellisimmät n-grammit mukaan lopulliseen malliin (Stolcke, 1998). Klusteroinnissa sekvenssien todennäköisyydet lasketaan sanojen sijaan sanojen muodostamien luokkien (esimerkiksi viikonpäivät tai ammattisanat) perusteella (Brown et al., 1990b).

2.4 Kielimallien sovelluksia

Eri sovellukset asettavat erilaisia vaatimuksia niin mallinnuksen tarkkuudelle ja laajuudelle kuin myös mallinnuksen yksikölle. Tässä työssä tarkastelun kohteeksi otetaan erityisesti kaksi sovellusta, automaattinen puheentunnistus ja konekäännös, jotka kuvataan seuraavassa. Mallinnuksen yksikön rooliin kussakin sovelluksessa palataan luvussa 4.

2.4.1 Puheentunnistus

Puheentunnistus on äänikäyttöisten käyttöliittymien perusta, mutta edellytys myös muun muassa puhetta sisältävästä aineistosta tehtävälle tiedonhaulle. Tehtävänä on muuntaa

puhuttu äänisignaali tekstiksi, jotta se voidaan tulkita tai sitä voidaan käsitellä automaattisesti. Toisin sanoen kohinaisen kanavan mallin kohinana ovat lausumisen piirteet ja tavoitteena on löytää todennäköisin vastaava kirjoituksena esitetty lause.

Puheentunnistuksessa on tapana erottaa toisistaan puhujasta riippumaton ja puhujariippuvainen tunnistus. Lisäksi puheen konteksti ohjaa mallinnusta – mikäli kyseessä on rajatun aihealueen tai jopa rajatun sanaston puhetta, sanasto on paljon pienempi kuin vapaamuotoisen puheen tapauksessa. Myös tunnettu puhuja, jonka lausumiseen akustinen malli opetetaan, helpottaa tunnistustehtävää.

2.4.2 Konekäännös

Kielenkäännös on monimutkainen tehtävä, jossa on otettava huomioon käännöksen konteksti ja tarkoitus, kirjoittajan tai puhujan tausta, tekstilajista riippuen sen ominaispiirteet ja niin edelleen. Etenkin kaunokirjallisuuden käännöksessä on välttämätöntä huomioida myös paljon sellaista, joka ei selviä edes teoksesta itsestään. Konekäännöksessä ongelmaa on yksinkertaistettu keskittymällä yhden tai useamman lauseen oikeaan käännökseen jättämällä tätä laajempi konteksti huomioimatta (Brown et al., 1990a).

Konekäännös voidaan ajatella puheentunnistusta vastaavana ongelmana, jossa akustisen signaalin sijaan lähtökohtana eli kohinaisena signaalina on vieras kieli. Käännökseen liittyy käännösmalli ja kohdekielen kielimalli. Käännösmalli rakentuu merkkijonoista, joille on laskettu todennäköisyys, että ne ovat käännös annetusta lähtökielen merkkijonosta (Jurafsky ja Martin, 2008).

Varsinainen kielimalli käsittelee pohjimmiltaan yleensä sanojen todennäköisyyksiä, joista lauseiden todennäköisyydet muodostuvat. Käännöstä mutkistaa se, ettei yksi sana tai lause alkuperäiskielessä välttämättä vastaa yhtä lausetta kohdekielessä. Lähtökielen sana voi tuottaa käännöksessä useamman sanan, tai kadota kokonaan.

Käännöksen erityishaasteena on, että sen opettaminen vaatii sopivan kaksikielisen aineiston. Tällaisen löytäminen on etenkin harvinaisemmille kielille vielä vaikeampaa kuin yksikielisen aineiston. Paljon käytettyjä aineistoja ovat esimerkiksi kaksikielisten valtioiden, kuten Suomi ja Kanada, sekä EU:n viralliset parlamenttitekstit ja lakitekstit, jotka ovat kahdella tai useamalla kielellä.

2.5 Kielimallien vertailukriteerit

Erilaisten mallien ja mallinnusratkaisuiden vertailu toisiinsa ei ole suoraviivaista. Ratkaisun toimivuus riippuu lopulta täysin sen käyttökontekstista ja sopivuudesta tehtäväänsä, mutta tutkimuksessa käytetään joitakin hieman yleisempiä vertailumetriikoita. Nämä luvut annetaan suhteessa johonkin vertailumalliin (*baseline*) ja opetus- ja testiaineistoihin,

joiden valinta edelleen mutkistaa eri lähteissä esitettyjen ratkaisujen vertailua toisiinsa. Vertailukriteerit voidaan jakaa kahteen luokkaan. Suora evaluaatio arvioi kielimallia itseään ja on sovellusriippumaton, kun taas epäsuora arvioi systeemiä kokonaisuutena (Jurafsky ja Martin, 2008).

Yleisin käytetty vertailuluku on informaatioteoriasta lainattu tunnusluku hämmentyneisyys (*perplexity*). Hämmentyneisyyden laskeminen on suora evaluaatio kielimallille, ja lukema voidaan tulkita niiden sanojen keskimääräiseksi lukumääräksi, joiden väliltä malli joutuu valitsemaan (Jurafsky ja Martin, 2008). Pienempi arvo tarkoittaa parempia ennustuksia tuottavaa mallia. Hämmentyneisyydellä voidaan myös arvioida kieltä – mitä korkeampi arvo, sitä monimutkaisempi kieli (Rosenfeld, 2000). Goodman (2001) huomauttaa kuitenkin, että vaikka hämmentyneisyys on paljon käytetty arviointikriteeri, se ei ole täysin luotettava. Hänen mukaansa alle 30% hämmentyneisyysvähennystä ei voi pitää luotettavasti tuloksen parantumisen merkkinä, ja sitä suuremmat parannukset ovat harvinaisia.

Toinen informaatioteoriasta periytyvä arvo on entropia (*entropy*), joka voidaan laskea hämmentyneisyydestä (tai toisin päin). Entropia on keskimääräinen vaadittu bittien lukumäärä sanaa kohti, joka tarvitaan testiaineiston kuvaamiseen optimaalisessa tapauksessa (Goodman, 2001).

Yleisin mitta epäsuorille evaluaatiomenetelmille on WER (*word error rate*). Se annetaan prosenttilukuna, joka on prosenttiarvo systeemin palauttaman sanajonon erolle oikeaan ratkaisuun (Jurafsky ja Martin, 2008). Morfologista tietoa hyödyntävien mallien arvioinnissa on käytetty myös LER (*letter error rate*) -arvoa, joka lasketaan merkeistä sanojen sijaan, sillä WER ei välttämättä sovellu taipuviin kieliin (Kurimo et al., 2006).

Konekäännöksessä käytetään myös Papineni et al. (2002) kehittämää BLEU-pisteytystä, jonka ajatuksena on automaattisesti verrata konekäännöksen tulosta ihmisten tuottamiin käännöksiin. Se perustuu WER-arvoon, mutta ottaa kuitenkin huomioon vaihtoehtoiset yhtä kelvolliset sanat ja sanajonot. Tuloksena on asteikko, joka korreloi vahvasti ihmisten tekimien arvioiden kanssa eri kielillä. Käännökselle annettu arvo on välillä 0–1, missä 1 merkitsee täydellistä vastaavuutta, mutta ihmiskäännöksetkin saavat tyypillisesti huomattavasti sitä alempia arvoja.

3 Mallinnusyksikkö

Seuraavassa käsitellään mallinnusyksikön roolia n-grammipohjaisessa kielimallissa ja esitellään mahdolliset yksiköt.

3.1 Yksikön rooli kielimallissa

Kielimallin yksikkö on pienin kielen osa, jota malli käsittelee. Sekvenssien todennäköisyydet muodostuvat n-grammimallissa yksikön ensiintymien todennäköisyyksistä. Instansseista kootaan sanasto (*lexicon*) – kutsumme sitä tässä sanastoksi riippumatta siitä, onko yksikkönä sana tai jokin muu. Sanaston kattavuus ja osuvuus ratkaisee mallin toimivuuden. Jos se on liian pieni tai koostuu vääristä sanoista, se ei kata mallinnettavaa kieltä tarpeeksi hyvin, eikä täten pysty antamaan siitä informaatiota. Liian suuri sanasto puolestaan on raskas käsitellä.

Kurimo et al. (2006) antavat puheentunnistuksessa käytetylle sanaa pienemmälle yksikölle neljä vaadittavaa ominaisuutta. Ensinnäkin saadun sanaston koon täytyy olla tarpeeksi pieni, että n-grammimallinnus on mielekkäämpää kuin sanoilla. Toiseksi, sanastosta rakennettavan kielen on oltava tarpeeksi kattava. Kolmanneksi yksiköiden pitää olla merkityksellisiä, jotta edellisiä instansseja voidaan käyttää seuraavien ennustuksessa n-grammimallissa. Viimeiseksi puheentunnistuksen tapauksessa yksikön lausuminen pitää voida määrätä.

Näistä kaikki paitsi viimeinen ehto voidaan yleistää vaatimuksiksi kielimallin yksiköllе sovelluksesta riippumatta. Yleisesti ottaen pitkät yksiköt vaativat suuremman sanaston, jotta siitä tulee kattava. Lyhyemmillä yksiköillä voidaan kattaa kieli pienemmällä sanastolla ja saada tarkempia estimaatteja opetusaineistosta, mutta vaaditaan pidempi n-grammi, jotta mallinettavasta kielestä voidaan laskea tarkkoja arvioita.

3.2 Mahdollisia yksiköitä

Seuraavassa esitellään erilaisia mahdollisia mallinnuksen yksiköitä ja kuvataan kunkin erityispiirteitä ja mallille asettamia vaatimuksia. Yksiköt ja esimerkkejä niistä suomen, ruotsin ja englannin kielistä on koottu taulukkoon 1.

Perinteisesti yleisin kielen mallinnuksessa käytetty yksikkö on sana (katso Brown et al., 1990a). Tämä johtuu muun muassa siitä, että useat paljon mallinnetuista valtakielistä – esimerkiksi englanti ja ranska – ovat suhteellisen analyyttisiä. Pitkät yhdyssanat ovat harvinaisia ja sanat itsenäisiä kielen yksiköitä. Lisäksi kirjoitusjärjestelmä erottelee sanat triviaalisti toisistaan. Tämä ei kuitenkaan pidä paikkaansa läheskään kaikille kielille, ja vaihtoehtoisten mallinnusyksiköiden tutkimus onkin pitkälti käynnistynyt muiden kielten tutkimuksen kautta.

Erityisesti agglutinatiivisten kielten mallinnukseen on kehitetty morfeemeihin perustuvia malleja. Nämä voidaan jakaa karkeasti kahteen luokkaan – kielitieteellisiin, ennalta määrättyihin morfeemeihin ja toisaalta automaattisesti opittuihin, ennalta määräämättömiin tilastollisiin morfeemeihin. Kielitieteelliset morfeemit vaativat niiden huolellisen määrit-

Taulukko 1: Kielen mallinnuksessa käytetyt yksiköt ja mahdollisia instansseja eri kielistä.

yksikkö	**suomi**	**englanti**	**ruotsi**
merkki	i	p	s
tavu	lä	prime	stor
morfeemi	ll[a]	er	st[o]r
morfi	llä	er	sta
sana	kelilläkään	prime	största
fraasi	liukas keli	prime minister	största bil

telyn etukäteen ja ovat väistämättä kieliriippuvaisia. Ne ovat kuitenkin intuitiivisesti houkuttelevia, sillä määrittelyllä voidaan varmistaa yksikön merkityksellisyys (Mihajlik et al., 2010). Tilastollisiin morfeemeihin liittyy oppimisalgoritmin kehittämisen ongelma, mutta koska sääntöjä ei tarvitse lyödä lukkoon, menetelmä on lähtökohtaisesti kieliriippumaton ja kerran kehitettyä menetelmää voidaan soveltaa kielestä toiseen (Kurimo et al., 2006).

Tilastolliset morfeemit eroavat kielitieteellisistä siinä, että niihin saattaa liittyä yli- tai alipilkontaa. Ylipilkonnassa sana pilkotaan liian pieniin palasiin, joilla ei välttämättä ole itsenäistä merkitystä – esimerkiksi sanan `minister` jako osiin `minist` + `er`. Alipilkonnassa useampi kieliopillinen morfeemi on liittynyt yhteen, esimerkiksi sanan `kelillä` käsittely yhtenä morfeemina.

Creutz (2006) vertaili kielellisiä ja tilastollisia morfeja puheentunnistuksen yksikkönä. Tilastolliset morfit pärjäsivät testeissä parhaiten, mutta kielelliset morfit tuottivat yhtä hyviä tuloksia aineistolla, jossa ei ole paljon lainasanoja ja erisnimiä. Mihajlik et al. (2010) raportoi parhaat tulokset yhdistelemällä tietoa kielellisistä ja tilastollisista morfeista. Kurimo et al. (2006) tutkivat puolestaan kielimallin kokoa suhteessa aineiston kokoon. Sanapohjaisessa mallissa erillisten sanojen määrä kasvaa lineaarisesti kun aineiston koko kasvaa, kun taas morfeilla se kasvaa hitaammin ja sitten vielä tasaantuu. Tämä osoittaa ainakin, että morfeemeilla voidaan rajoittaa sanaston kokoa ja saada pienemmällä aineistolla kattavampi malli.

Epätriviaalia yksikköä kuten sanaa pienempää yksikköä käytettäessä on myös määriteltävä, miten luonnollinen kieli voi rakentua näistä palasista – esimerkiksi miten morfeemeista voidaan rakentaa sanoja. Tässä joudutaan tekemään oletuksia sanojen pituudesta ja rakenteesta, mikä heikentää ainakin teoriassa hieman tilastollistenkin morfien yleistettävyyttä eri kieliin. Toisissa malleissa sanat jaetaan vain sanavartanloon ja loppuosaan (Sarikaya et al., 2008), joka voi olla taivutus- tai sijamuotopääte tai niiden yhdistelmä. Toisissa päätteitä tai etuliitteitä voi olla vapaa määrä (Creutz, 2006).

Vielä pienempiä yksiköitä ovat tavut ja yksittäiset merkit. Kielestä riippuen yksi merkki saattaa sisältää hyvin vähän tai hyvin paljon informaatiota. Esimerkiksi suomen kielen kohdalla merkki `i` ei yksinään vielä merkitse juuri mitään tai kerro siitä, millaisia merkkejä sitä ennen voi esiintyä. Sen sijaan kiinan merkki on jo itsessään sana tai morfeemi. Käytännössä puhtaasti merkkipohjaista mallinnusta ei olla juurikaan tehty. Kiinan kaltaisten kielten kohdalla kuitenkin sana, tavu, merkki ja morfeemi ovat hyvin lähellä toisiaan.

Tavu on lähtökohtaisesti äänteellinen yksikkö, joten sen sovittaminen puheentunnistuksessa akustisen mallin kanssa on vaivattomampaa kuin esimerkiksi morfeemin. Joissakin kielissä, kuten kiinassa, tavu sisältää jo jonkin verran informaatiota, kun taas fonologisesti köyhissä kielissä kuten japanissa yksi tavu on erittäin monimerkityksinen. Thaissa tavu vastaa melko usein yhtä sanaa, ja sitä ollaankin mallinnettu käyttäen yksikkönä tavua (Wutiwiwatchai ja Furui, 2007).

Sanoja isompia yksiköitä ovat fraasit ja kollokaatiot. Kollokaatiolla tarkoitetaan karkeasti kahta tai useampaa sanaa, jolla on taipumusta esiintyä yhdessä tietyssä järjestyksessä (Jurafsky ja Martin, 2008), ja jotka näin tyypillisesti muodostavat yhdessä merkitysyksikön. Esimerkiksi erisnimet, paikannimet ja totutut sanonnat ovat kollokaaitoita. Fraaseja on käytetty etenkin konekäännöksessä (Federico ja Bertoldi, 2005), mutta myös puheentunnistuksen kielimallissa (Kuo ja Reichl, 1999).

3.3 Yksiköitä yhdistelevät mallit

Mallinnus voidaan tehdä myös useammalla tasolla, jolloin pyritään saamaan etu useammasta yksiköstä. Tyypillisintä on säilyttää sanaesitys, mutta tallentaa myös sen jako morfeemeihin. Tällöin kaikkein tavallisimmat sanat ovat mallissa laskennallisesti nopeammin saatavilla, kuin ne olisivat jos ne pitäisi rakentaa morfeemeista. Lisäksi sanojen kokoaminen uudelleen on helpompaa, kun rakenne on säilytetty. Hierarkista mallinnusta ovat tutkineet muun muassa Sarikaya et al. (2008) ja Niessen ja Ney (2004). Chiang (2005) käyttää hierarkista fraasimallia konekäännöksen käännösmallissa.

Sarikaya et al. (2008) rakensi arabian morfeemeista ja sanoista puumallin, joka paransi mallinnustuloksia sekä puheentunnistuksessa että konekäännöksessä. Niessen ja Ney (2004) mallinsivat konekäännöstä varten saksaa säilyttämällä sanan perusmuodon ja luokkia kuten sanaluokka ja sijamuoto.

Samaan tapaan faktorimallit (*factored language models*) käyttävät yksikkönään sanaa, joka on esitetty erilaisten piirteiden vektorina (Bilmes ja Kirchhoff, 2003). Piirteinä voi toimia muun muassa morfologisia luokkia, sanan perusmuotoja tai sanaluokkia. Piirteet mahdollistavat sen, ettei n-grammin pituutta tarvitse kasvattaa kuten sanaa pienempää yksikköä käytettäessä, mutta muun muassa morfologista tietoa voidaan silti sisällyttää

kielimalliin. Bilmes ja Kirchhoff (2003) saavuttivat käyttäen faktorimallia ja siihen sopivaa smoothing-tekniikkaa 2-grammeilla 3-grammejakin paremmat tulokset englannilla ja arabialla.

4 Yksikön valintaan vaikuttavat seikat

Tässä osiossa eritellään yksikön valintaan vaikuttavia seikkoja ja tarkastellaan yksitellen niiden kunkin vaikutusta yksikön valintaan ja mallinnuksen lopputulokseen.

4.1 Kieli

Erilaisten kielten rakenteessa ja esitystavoissa on ennalta-arvaamattomia eroja, joita on vaikea sovittaa yhteen malliin. Useimmissa sovelluksissa tiedetään mallinnettava kieli tai kielet etukäteen, ja tämä voidaan ottaa tarvittaessa huomioon. Monet tutkijat keskittyvätkin etsimään parhaita ratkaisuja tiettyjen kielten erityispiirteiden kautta (katso Ablimit et al., 2010; Kudo et al., 2004; Wutiwiwatchai ja Furui, 2007).

Toisaalta kieliriippumattomat mallit olisivat toivottavia, sillä ne säästäisivät rinnakkaista kehitystyötä. Puhtaasti tilastollisilla menetelmillä on saavutettu lupaavia tuloksia sekä puheentunnistuksessa että konekäännöksessä (katso Creutz, 2006; Mihajlik et al., 2010; Federico ja Bertoldi, 2005). Käsiteltävän aineiston tyypistä riippuen on myös otettava huomioon, että luonnollisen kielen seassa on usein myös useampaa kuin yhtä kieltä lainasanojen, erisnimien ja viittauksien muodossa. Näissä tapauksissa kielestä riippumattomat ratkaisut toimivat parhaiten.

Mallinnuksen kannalta merkittäviä kielten eroja ovat triviaalin sanajaon olemassaolo, morfien keskimääräinen lukumäärä sanaa kohden, morfien erottuminen toisistaan, sanojen ja lauseiden rakentuminen morfeista sekä puhutun ja kirjoitetun kielen vastaavuus. Kielestä riippuen osa mahdollisista yksiköistä, kuten tavu ja merkki, merkki ja morfi tai morfi ja sana voivat vastata toisiaan. Se, mikä toisessa kielessä ilmaistaan yhdellä sanalla saattaa toisessa vaatia kokonaisen fraasin tai lauseen.

Nämä eri yksiköt ovat kuitenkin teoreettisia käsitteitä, joten puhtaasti tilastollista yksikköä käyttävän mallin tapauksessa erilaisten yksiköiden roolien vaihtelu ei väistämättä ole ongelma. Kielestä voidaan aina erottaa toistuvia palasia, mutta sanojen ja lauseiden kokoaminen niistä vaatii oletuksia siitä, miten kieli rakentuu.

Kirjoituksessa triviaalin yksikön olemassaolo riippuu kielen kirjoitusjärjestelmästä. Muun muassa latinalaisin ja kyrillisin aakkosin kirjoitetuissa kielissä sana on selkeä vaihtoehto yksiköksi, sillä se erottuu välein ja välimerkein. Tämä ei kuitenkaan pidä paikkaansa monille Aasian kielille kuten kiina ja thai. Japanin kirjoituksessa käytetään sekä Kiinas-

ta peräisin olevia symbolimerkkejä että tavumerkkejä, joten sanarajojen löytäminen on kiinaa helpompaa limittäisten kirjoitusjärjestelmien vuoksi. Jos triviaalia yksikköä ei ole, morfologista analyysiä joudutaan tekemään joka tapauksessa, ja sanarajojen löytäminen voidaan rinnastaa morfeemeihin jaon ongelmaan (Kudo et al., 2004).

Synteettisissä kielissä kuten suomi ja turkki, joissa sanat koostuvat useista morfeista, sana ei ole mielekäs yksikkö, sillä sanasto kasvaa tällöin liian suureksi (Sarikaya et al., 2008; Hirsimäki et al., 2009). Näillä kielillä morfeihin perustuvilla menetelmillä on saatu hyviä tuloksia (Mihajlik et al., 2010; Kurimo et al., 2006). Yhdyssanoja paljon käyttävien kielten kuten saksan kohdalla morfeemijaon hyödyllisyydestä puolestaan on ristiriitaisia tuloksia (Niessen ja Ney, 2004; Larson et al., 2000). Annetuissa esimerkeissä on kuitenkin käytetty eri menetelmiä, joten niiden vertailu on vaikeaa. Myös Yang ja Kirchhoff (2006) huomasivat, että sanojen jakamisesta on konekäännöksessä enemmän hyötyä suomelle kuin saksalle.

Myös morfeemijaon vaativuudessa on eroa kielten välillä. Arabia on esimerkki kielestä, jossa kaikki morfit eivät erotu toisistaan yksinkertaisella pilkkomisella. Siinä sanan vartalo koostuu konsonanttirungosta ja siihen liitettävistä vokaaleista – esimerkiksi vartalosta `k-t-b` voidaan muodostaa sanat `kitaab` (kirja) ja `kaatib` (kirjoittaja) (Sarikaya et al., 2008). Tähän vartaloon voi myös liittyä eri määrä prefiksejä ja suffikseja. Sarikaya et al. (2008) saivat parhaat tulokset arabiassa rajoittamalla mahdollisten etuliitteiden ja suffiksien määrän yhteen jokaista sanaa kohden. Hyvin fuusioivassa slovenian kielessä ei puolestaan saatu yksiselitteisiä parannuksia käyttämällä sanan sijasta lyhyempää yksikköä (Maucec et al., 2009).

4.2 Sovellus

Kielimallin hyvyys riippuu lopulta siitä, mihin sillä pyritään. Vaikka samoja mallinnuskeinoja on onnistuneesti käytetty eri sovelluksissakin, sovellus on silti keskeinen mallinnuksen vaatimukset määrittelevä tekijä. Sovelluskohtaisia haasteita mallinnuksen yksikölle käsitellään tässä konekäännöksessä ja puheentunnistuksessa.

4.2.1 Puheentunnistus

Puheentunnistuksen erityispiirteenä on se, että kielimallin on toimittava yhdessä akustisen mallin kanssa. Tällöin kielimallin yksiköstä on voitava luoda yhteys ääntämykseen. Suomensukuisissa kielissä pienestäkin yksiköstä on yleensä helppoa konstruoida oikea äänneasu, kun taas esimerkiksi englannin kohdalla se on huomattavasti vaikeampaa (Kurimo et al., 2006). Hirsimäki et al. (2009) osoittivat, että n-grammin pituutta kasvattamalla voidaan ratkaista lyhyeen yksikköön liittyviä ongelmia, mutta he käyttivät kokeessaan

vain suomen ja viron kieliä.

Parhaat tulokset puheentunnistuksessa riippuvatkin pitkälti kielestä – agglutinatiivisille kielille ne on saatu morfipohjaisin menetelmin (muun muassa Maucec et al., 2009; Hirsimäki et al., 2009), äärimmäisen isoloiville tavuin (Wutiwiwatchai ja Furui, 2007). Sanaston koon kannalta on syytä erotella suljetun ja avoimen sanaston puheentunnistus. Suljetusta sanastosta voidaan puhua esimerkiksi silloin, kun kyseessä on hyvin tarkasti rajatun aihealueen sovellus. Tällöin sanaston kasvu taipuvillakaan kielillä ei ole vastaava kuin avoimen sanaston puheentunnistuksessa. Hyvin rajatun aihealueen puheentunnistuksessa myös fraasit saattavat toimia (Kuo ja Reichl, 1999), sillä hyvin määritellyssä tilanteessa kuten varusjärjestelmässä samoilla fraaseilla on taipumusta toistua.

4.2.2 Konekäännös

Konekäännöksessä kielimallin tehtävä on tuottaa mahdollisimman sujuvaa kohdekieltä. Toisin kuin puheentunnistuksessa, siinä liikutaan laajemmalla merkitystasolla, jossa esimerkiksi kollokaatiot ovat tärkeitä. Konekäännöksessä on usein keskitytty kääntämään vieraista kielistä englantiin, mutta paljon haastavampi tehtävä on tuottaa morfologisesti rikasta kohdekieltä, sillä lähtökieli ei sisällä eksplisiittisesti kaikkea tarvittavaa informaatiota (Virpioja et al., 2007; Minkov ja Toutanova, 2007). Tästä huolimatta Kirchhoff ja Yang (2005) tulivat siihen tulokseen, että kielimalli voi parantaa konekäännöksen tulosta vain vähän, ellei myös käännösmallia paranneta.

Konekäännöksessä käytössä on aina kaksikielinen aineisto, joka mahdollistaa mallin rakentamisen molempien kielten piirteet huomioon ottaen. Ma ja Way (2009) osoittivat että käännöksen tuloksia voidaan parantaa tekemällä sanaa pienempiin yksiköihin jako hyödyntäen kaksikielistä aineistoa.

Käännösmallissa fraasit ja hierarkinen mallinnus ovat osoittautuneet toimiviksi ratkaisuiksi (Niessen ja Ney, 2004; Chiang, 2005). Tähän on syynä osin se, että käännöksessä myös sanojen järjestys usein muuttuu kielten välillä, joten toisiinsa liittyvien sanojen suhteiden ottaminen mukaan malliin parantaa tuloksia. Myös puheentunnistuksessa käytettyn tilastollisten morfeemien menetelmän yhdistäminen sanamalliin konekäännöksessä tuotti parannuksia (de Gispert et al., 2009). Toisin sanoen vaikuttaa siltä, että konekäännöksen tapauksessa joudutaan ottamaan huomioon puheentunnistusta pidempi konteksti.

4.3 Muut tekijät

Edellä käsiteltyjen seikkojen lisäksi erilaiset käytännön rajoitteet voivat ohjata yksikön valintaa. Mitä lyhympi yksikkö on, sitä pienempi on sanasto, mutta sitä pidempiä n-grammeja tarvitaan, kun taas pitkä yksikkö vaatii ison sanastoa ja paljon opetusaineistoa.

Täten käytettävissä oleva aineisto voi ohjata yksikön valintaa etenkin tilanteissä, joissa saatavilla oleva aineisto on rajallinen. Tällaisia tilanteita ovat esimerkiksi, kun halutaan mallintaa harvinaista kieltä tai kielen genreä kuten puhekieltä. Kaikkein harvinaisimpien kielten sekä myös puhekielen kohdalla myös kielitieteellisiä morfeemeja varten tarvittava kielitieteellinen määrittely voi puuttua tai olla liian teoreettinen.

Mallinnettava kieli voidaan joissakin tilanteissa myös olettaa yksinkertaisemmaksi kuin mitä se on yleisessä tapauksessa. Esimerkiksi hyvin suljetun alan kielen kohdalla, kuten vaikkapa säätiedotuksissa, sanasto on hyvin rajattu ilman morfologistakin käsittelyä ja täten sana voi olla hyvä yksikkö synteettisessäkin kielessä. Hyvin kaavamuotoisessa kielessä fraasi saattaa olla paras ratkaisu jopa puheentunnistuksessa (Kuo ja Reichl, 1999). Lisäksi suomen kielessä puhekieli on huomattavasti yleiskieltä isoloivampaa, joten sen haasteet voivat olla toisaalla. Tätä ei ole kuitenkaan ole vielä tutkittu.

Toisaalta toisinaan voidaan erityisesti tarvita kieliriippumaton lähestymistapa. Tällaisia tilanteita ovat esimerkiksi, kun mallinnettavassa kielessä esiintyy paljon erisnimiä ja lainasanoja. Creutz (2006) huomasi, että tilastolliset morfit pärjäsivät kieliopillisia morfeja paremmin etenkin uutisaineistossa, jossa vieraskielisiä sanoja ja nimiä on paljon.

5 Tulokset

Edellä käsiteltiin yksitellen yksikön valintaan vaikuttavia päätekijöitä – kieltä, sovellusta ja muita tekijöitä kuten käytössä olevia resursseja. Mallinnusyksikkö toimii eri malleissa ei tavoin, mutta eri malleja ei ole nimenomaan yksikön valinnan kannalta juurikaan vertailtu keskenään. Tässä luvussa tarkastellaan tämänhetkisen tutkimuksen valossa yksikön valintaa kokonaiskysymyksenä ja pyritään osoittamaan kaikkein merkitsevimmät yksikön valintaan vaikuttavat tekijät.

5.1 Merkittävimmät valintakriteerit

Mallinnuksen yksikkö voidaan valita joko pyrkimällä parhaaseen mahdolliseen tulokseen kaikki eri tekijät huomioon ottaen, tai pyrkimällä mahdollisimman yleiseen menetelmään. Saman menetelmän käyttö kielestä tai sovelluksesta toiseen säästää vaivaa ja kustannuksia. Perimmiltään joudutaan punnitsemaan yleistyksen tuoma etu suhteessa kontekstia varten optimoituun ratkaisuihin. Esimerkiksi hyvin harvinaisen kielen tai lyhytikäisen sovelluksen tapauksessa saattaa olla parempi tyytyä johonkin helposti toteutettavaan, yleisesti toimivaksi todettuun ratkaisuun.

Kuten edellä esitettiin, mallinnettava kieli vaikuttaa väistämättä kielimalliin ja yksikön valintaan, sillä eri kielissä eri yksiköiden merkitys on hyvin erilainen. Eri kielissä parhaat

tulokset onkin tällä hetkellä raportoitu eri mallinnusyksiköitä käyttäen. Tämä johtuu osin kielten erityispiirteistä, mutta voi osin olla seurausta myös tutkimuksen keskittymisestä kielikohtaisiin ratkaisuihin.

Mallinnuksen parhaan yksikön määräävä tekijä ei vaikuttaisi olevan tietty kieli eikä edes kielikunta, vaan kaksi vielä yleisempää piirrettä – agglutinatiivisuuden ja fuusioitumisen aste. Esimerkiksi turkkia ja suomea, jotka ovat molemmat agglutinatiivisia mutta eivät mitään sukua toisilleen, on onnistuneesti mallinnettu samalla tilastolliseen morfiin pohjautuvalla menetelmällä (Kurimo et al., 2006). Sen sijaan siirryttäessä morfologiselta rakenteeltaan erilaisempiin kieliin yleistäminen vaikuttaisi käyvän hankalammaksi. Tämä voi olla syynä esimerkiksi siihen, miksi agglutinatiivisissa kielissä hyviksi koetut menetelmät eivät ole toimineet aivan yhtä hyvin englannissa (Creutz, 2006) tai saksassa (Larson et al., 2000).

Kuitenkin myös hyvinkin etäisten kieliparien kuten arabia ja venäjä tai suomi ja englanti, kohdalla ollaan osoitettu lupaavia tuloksia erityisesti tilastollista morfia käyttäen (Creutz, 2006; Minkov ja Toutanova, 2007). Vielä on selvittämättä, kuinka hyvin tämä menetelmä on yleistettävissä vielä kaukaisempiin kieliin.

Sovelluksessa yhteistä määrittävää tekijää on vaikeampi osoittaa, sillä tulokset vaihtelevat tarkemman sovelluskontekstin mukaan. Puheentunnistuksessa toimivin menetelmä riippuu edellä esitettyjen tulosten valossa vahvasti kielestä, aineistosta ja sanaston koosta. Konekäännöksessä sen sijaan kielimallin rooli on vähemmän tärkeä, ja käännösmallissa eri yksiköitä yhdistelevät ratkaisut ovat parhaita (Niessen ja Ney, 2004; Yang ja Kirchhoff, 2006; Chiang, 2005). Tähän syynä voidaan nähdä käännöstehtävän lähtökohtainen monitasoisuus tai se, että käännökseen vaikutta aina kaksi kieltä ja yhdistely mahdollistaa eri kieliin yleistämisen.

Muita valintaan vaikuttavia tekijöitä ovat saatavilla olevan aineiston määrä ja mallinnettavan kielen genre ja mahdollinen rajaus. Sanaa pienemmällä yksiköllä voidaan käyttää pienempää aineistoa (Niessen ja Ney, 2004; Maucec et al., 2009; Ablimit et al., 2010), kun taas sanamallilla on raportoitu parempia tuloksia suuremmilla aineistoilla. Vaadittava laskentateho sen sijaan nähdään joko toissijaisena kysymyksenä tutkimuksen kannalta, kuten Goodman (2001) huomautti, tai luotetaan siihen että saatavilla oleva laskentateho jatkaa eksponentiaalista kasvuaan.

5.2 Lupaavat yksiköt

Kun puheentunnistuksessa halutaan valita paras mallinnusyksikkö mallinnettavan kielen mukaan, paras yksittäinen yksikkö vaikuttaisi kielen synteettisyyden asteesta riippuen olevan tavu, sana tai tilastollinen morfi. Tilastollisen morfin tehokkuus on osoitettu usealla eri agglutinatiivisella kielellä (Creutz, 2006; Sarikaya et al., 2008; Mihajlik et al.,

2010; Kurimo et al., 2006). Esimerkiksi Larson et al. (2000) saivat parempia tuloksia sanamallilla saksan tapauksessa, vaikka saksassakin on sijamuotoja ja pitkiä yhdyssanoja.

Konekäännöksessä parhaasta yksiköstä ei ole ehdottomia tuloksia, mutta se vaikuttaisi olevan astetta puheentunnistuksen ideaaliyksikköä pidempi. Sanapohjaisella mallilla on saatu hyviä tuloksia kahden analyyttisen kielen tapauksessa (Mariño et al., 2006) ja fraasipohjaisilla analyyttisissä ja isoloivissa kielissä (Federico ja Bertoldi, 2005).

Kieliriippumattomassa mallinnuksessa kielestä riippuvat käsitteet kuten sana joudutaan hylkäämään. Pyrittässä mahdollisimman yleistettävään, eri kielissä ja sovelluksissa toimivaan ratkaisuun, lupaavia lähestymistapoja on kaksi. Ensimmäinen niistä on tilastollinen morfi, joka on tähän asti osoitettu kaikkein yleistettävämmäksi yksiköksi, sillä se ei tee juurikaan oletuksia mallinnettavan kielen rakenteesta.

Toinen yleistämiseen sopiva ratkaisu on eri yksiköiden yhdistely. Tämä voidaan tehdä rakentamalla hierarkisia malleja, faktorimalleja tai yksinkertaisesti yhdistelemällä kahta eri yksikköä käyttävän mallin tulokset (de Gispert et al., 2009). Puheentunnistuksessa Maucec et al. (2009) osoittivat että morfi- ja sanaesityksellä on kummallakin etunsa, sillä morfit vähentävät tuntemattomien sanojen määrää mutta eivät mallin hämmentyneisyyttä, jolla mitattuna parhaat tulokset saatiin kuitenkin sanoilla. Yang ja Kirchhoff (2006) esittävät, että opetusaineistosta löytyviä ja tuntemattomia sanoja tulisi käsitellä eri tavoin, jolloin tunnetuille sanoille saadaan sanamallin edut mutta tuntemattomat sanat voidaan käsitellä morfitasolla. Myös de Gispert et al. (2009) onnistuivat parantamaan sanapohjaisen konekäännöksen tuloksia yhdistelemällä siihen puheentunnistukseen kehitetyn tilastollisen morfin, joka ei kuitenkaan yksin parantanut käännöstuloksia (Virpioja et al., 2007).

6 Yhteenveto

Tämän työn tavoitteena oli kartoittaa mallinnusyksikön roolia erityisesti n-grammipohjaisissa kielimalleissa sekä selvittää yksilön valintaan vaikuttavia seikkoja. Esimerkkisovelluksina käytettiin puheentunnistusta ja konekäännöstä. Työssä tutkittiin erilaisia yksiköitä eri kielissä ja kielten eroavaisuuksia mallinnusyksiköiden suhteen. Lisäksi käsiteltiin esimerkkisovellusten sekä kulloinkin käytössä olevien resurssien kielimallin yksikölle asettamia vaatimuksia.

Yksikön valinta vaikuttaa kattavan kielimallin rakentamiseen vaadittavan aineiston kokoon sekä mallin itsensä kokoon. Lyhyempi yksikkö mahdollistaa pienemmän aineiston käyttämisen, mutta vaatii pidemmän n-grammin ollakseen merkityksellinen. Onnistuneesti käytettyjä yksiköitä ovat tavu, morfeemi, kieliopillinen ja tilastollinen morfi sekä fraasi. Myös useiden eri yksiköiden piirteitä yhdistelevät ratkaisut ovat tuottaneet hyvä tulok-

sia. Paras yksikkö riippuu resurssien lisäksi kuitenkin myös kielestä ja jossakin määrin myös sovelluksesta.

Vaikka samaa yksikköä ja pilkontamenetelmää on onnistuneesti käytetty myös eri kielissä ja sovelluksissa, on yksikön valinta silti kysymys, johon ei ole yhtä parasta ratkaisua. Kaikkein triviaaleimman yksikön, joka useimmissa kielissä on sana, käyttäminen ei useinkaan ole paras vaihtoehto. Kieli on myös tilastollisessa kielen mallinnuksessa ymmärrettävä systeeminä, joka koostuu eri kokoisista merkitysyksiköistä.

Tämän työn tulokset pätevät vain n-grammimalleihin. Vaikka se onkin tyypillisin mallinnustekniikka, saattaa yksikön rooli olla aivan erilainen eri tekniikoissa. Morfologista tietoa voidaan lisäksi integroida mallinnukseen myös muuta kautta kuin suoraan n-grammin yksikkönä, esimerkiksi kokonaan kielimallin ulkopuolisena tiedonlähteenä (Sak et al., 2009).

Sovelluksen ja parhaan yksikön suhteesta ei saatu selkeää vastausta. Vaikuttaisi siltä, että sovelluksen tarkempi konteksti, kuten mallinnettavan kielen genre ja oletetun sanaston koko ovat merkitseviä tekijöitä yksikön valinnan kannalta. Eri sovellustilanteista ei kuitenkaan vielä ole juurikaan vertailutuloksia. Olisi esimerkiksi mielenkiintoista tietää, vaatiiko suomen yleiskieltä morfologisesti yksinkertaisempi puhekieli yhtä lyhyen yksikön puheentunnistuksessa. Lisäksi eri yksiköillä saadun tiedon yhdistelyn – muun muassa hierarkisten mallien ja faktorimallien – hyödystä ja yleistettävyydestä tarvitaan vielä lisää vertailutuloksia.

Erilaisten kielimallien vertailu toisiinsa on vaikeaa, koska niihin itseensä liittyy paljon erilaisia teknisiä ja sovelluskontekstiin liittyviä ratkaisuja ja myös arviointikriteerit eivät välttämättä ole yhteensopivia. Siksi myös yksikön vaikutuksen vertailu suhteessa erilaisiin muuttujiin on hankalaa. Tässä työssä onnistuttiin kuitenkin erittelemään yksikön valintaan vaikuttavia tekijöitä ja osoittamaan, kuinka eri tavoin eri yksiköt käyttäytyvät eri kielissä ja sovelluskonteksteissa.

Lähteet

Mijit Ablimit, Graham Neubig, Masato Mimura, Shinsuke Mori, Tatsuya Kawahara ja Askar Hamdulla. Uyghur morpheme-based language models and ASR. *IEEE 10th International Conference on Signal Processing (ICSP)*, Beijing, 10 2010.

Jeff Bilmes ja Katrin Kirchhoff. Factored language models and generalized parallel backoff. *Human Language Technology Conference/North American chapter of the Association for Computational Linguistics (HLT/NAACL-2003)*, Edmonton, Alberta, May/-June 2003.

Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer ja Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, sivut 79–85, 1990a.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai ja Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18:18–4, 1990b.

Stanley F. Chen ja Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393, 1999.

David Chiang. A hierarchical phrase-based model for statistical machine translation. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, sivut 263–270, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

Mathias Creutz. *Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition.* Väitöskirja, Teknillinen korkeakoulu, Espoo, 2006.

Adrià de Gispert, Sami Virpioja, Mikko Kurimo ja William J. Byrne. Minimum bayes risk combination of translation hypotheses from alternative morphological decompositions. *HLT-NAACL (Short Papers)*, sivut 73–76. The Association for Computational Linguistics, 2009.

Marcello Federico ja Nicola Bertoldi. A word-to-phrase statistical translation model. *ACM Trans. Speech Lang. Process.*, 2:1–24, December 2005.

Joshua Goodman. A bit of progress in language modeling. Tekninen raportti, Microsoft, Redmond, WA, 2001.

Teemu Hirsimäki, Janne Pylkkönen ja Mikko Kurimo. Importance of high-order n-gram models in morph-based speech recognition. *IEEE Transactions on Audio, Speech & Language Processing*, 17(4):724–732, 2009.

Daniel Jurafsky ja James H. Martin. *Speech and Language Processing (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall, Upper Saddle River, NJ, toinen painos, 2008. ISBN 0131873210.

Fred Karlsson. *Yleinen kielitiede*. Gaudeamus Helsinki University Press, Helsinki, 2008. ISBN 9789524950718.

Katrin Kirchhoff ja Mei Yang. Improved language modeling for statistical machine translation. *Proceedings of the ACL Workshop on Building and Using Parallel Texts ParaText 05*, sivut 125–128. Association for Computational Linguistics, 2005.

Taku Kudo, Kaoru Yamamoto ja Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. *In Proc. of EMNLP*, sivut 230–237, 2004.

Hong-Kwang Jeff Kuo ja Wolfgang Reichl. Phrase-based language models for speech recognition. *EUROSPEECH*, 1999.

Mikko Kurimo, Antti Puurula, Ebru Arisoy, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Tanel Alumäe ja Murat Saraclar. Unlimited vocabulary speech recognition for agglutinative languages. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, sivut 487–494, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Martha Larson, Daniel Willett, Joachim Köhler ja Gerhard Rigoll. Compound splitting and lexical unit recombination for improved performance of a speech recognition system for german parliamentary speeches. *INTERSPEECH*, sivut 945–948. ISCA, 2000.

Yanjun Ma ja Andy Way. Bilingually motivated domain-adapted word segmentation for Statistical Machine Translation. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, sivut 549–557, Athens, Greece, 2009.

José B. Mariño, Rafael E. Banchs, Josep Maria Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa ja Marta R. Costa-Jussà. *N*-gram-based machine translation. *Computational Linguistics*, 32(4):527–549, 2006.

Mirjam Sepesy Maucec, Tomaz Rotovnik, Zdravko Kacic ja Janez Brest. Using data-driven subword units in language model of highly inflective slovenian language. *IJPRAI*, sivut 287–312, 2009.

Péter Mihajlik, Zoltán Tüske, Balázs Tarján, Bottyán Németh ja Tibor Fegyó. Improved recognition of spontaneous hungarian speech: morphological and acoustic modeling

techniques for a less resourced task. *Trans. Audio, Speech and Lang. Proc.*, 18:1588–1600, August 2010.

Einat Minkov ja Kristina Toutanova. Generating complex morphology for machine translation. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, sivut 128–135, 2007.

Sonja Niessen ja Hermann Ney. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204, 2004.

Kishore Papineni, Salim Roukos, Todd Ward ja Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, sivut 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

Ronald Rosenfeld. Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE*, osa 88, sivut 1270–1278, 2000.

Hasim Sak, Murat Saraclar ja Tunga Gungor. Integrating morphology into automatic speech recognition. *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, 2009.

Ruhi Sarikaya, Mohamed Afify, Yonggang Deng, Hakan Erdogan ja Yuqing Gao. Joint morphological-lexical language modeling for processing morphologically rich languages with application to dialectal arabic. *IEEE Transactions on Audio, Speech and Language Processing*, 16(7):1330–1339, 2008.

Andreas Stolcke. Entropy-based pruning of backoff language models. *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, sivut 8–11, 1998.

Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz ja Markus Sadeniemi. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. *Proceedings of the Machine Translation Summit XI*, sivut 491–498, Copenhagen, Denmark, September 2007.

Chai Wutiwiwatchai ja Sadaoki Furui. Thai speech processing technology: A review. *Speech Communication*, 49(1):8 – 27, 2007.

Mei Yang ja Katrin Kirchhoff. Phrase-based backoff models for machine translation of highly inflected languages. *in Proceedings of the 21st International Conference on Computational Linguistics*, sivut 1017–1020, 2006.