TURUN YLIOPISTO

Taligenkänning av finlandssvenska och sverigesvenska

Utmaningar vid pluricentriska språk

Elisa Liipo
Avhandling pro gradu
Språkexpertprogrammet, Nordiska språk
Institutionen för språk- och översättningsvetenskap
Humanistiska fakulteten
Åbo universitet
Juni 2024

Avhandling pro gradu

Språkexpertprogrammet, Nordiska språk
Elisa Liipo
Taligenkänning av finlandssvenska och sverigesvenska. Utmaningar vid pluricentriska språk.
Antal sidor: 61 s., 5 s. bilagor

Denna avhandling behandlar pluricentriska språk ur taligenkänningssynvinkel. Svenska är ett pluricentriskt språk med två varianter, finlandssvenska och sverigesvenska. Syftet med min avhandling är att utreda hur ett taligenkänningsverktyg tolkar de två språkvarianterna, varför eventuella skillnader uppstår och i fall taligenkänningen gör samma feltolkningar vid båda språkvarianterna. Slutligen för jag en diskussion om hur taligenkänning kan vidareutvecklas.

Mitt material består av ljudklipp på finlandssvenska och sverigesvenska och taligenkännings tolkning av talet. Ljudklippen innehåller politiska debatter och diskussioner. Det finlandssvenska materialet består av 10 932 ord vilket motsvarar 88 minuter totalt och det sverigesvenska materialet består av 11 046 ord vilket motsvarar 68 minuter. Studien är både kvantitativ och kvalitativ till sin natur och jag kombinerar WER (*word error rate*) som metod med komparativa metoder.

Resultaten visar att det finns en tydlig skillnad i hur taligenkänning behandlar finlandssvenska och sverigesvenska. Den genomsnittliga WER-felprocenten var 0,14 % för finlandssvenska och 0,07 % för sverigesvenska. Det verkar som att den största faktorn som påverkar resultaten är mängden träningsdata. Eftersom taligenkänning kan användas som hjälpmedel av t.ex. äldre personer eller personer med funktionsnedsättning, måste detta uppmärksammas. Vi måste hitta lösningar för hur man kan insamla mer och bättre träningsdata även för icke-dominanta språkvarianter av pluricentriska språk. Detta är viktigt om vi ska kunna arbeta tillsammans för att bygga ett mer tillgängligt samhälle för alla dess invånare.

Ämnesord: taligenkänning, språkteknologi, finlandssvenska, sverigesvenska

Innehåll

1 Inledning

Språkteknologi är ett tvärvetenskapligt område där språkvetenskap möter datavetenskap. Under de senaste åren har språkteknologi blivit ett allt större fält och utvecklingen av olika modeller och verktyg har varit betydande. Språkteknologi syftar till att utveckla och tillämpa datorbaserade system och verktyg för att analysera, förstå och generera naturligt språk. För att skapa bra språkteknologiverktyg krävs det stora mängder språkdata som används för att lära datorprogram bearbeta naturligt språk. Materialet som utgör träningsdata är givetvis lättare att samla in för ett språk som talas av majoriteten i ett land, till exempel finska i Finland respektive med svenska i Sverige.

Taligenkänning är en av de viktigaste tillämpningarna inom språkteknologi och fokuserar på att konvertera tal, alltså ljud, till skriftlig form (Jurafsky och James 2023). Ett exempel på applikationer som utnyttjar taligenkänning är den välkända digitala assistenten *Siri*. En av utmaningarna inom taligenkänning är regionala och dialektala varieteters uttal. När taligenkänningsverktyg tränas med data från en viss geografisk region eller en specifik standardvariant kan verktyget ha svårigheter att korrekt tolka talare som talar en annan regional variant.

Pluricentriska språk är språk som har en officiell status i fler än ett land (Clyne 1992). Detta innebär att pluricentriska språk kan ha betydande skillnader i uttalsmönster mellan olika länder och regioner där språket talas. Det här gör pluricentriska språk såväl till ett intressant som utmanande ämne att studera inom taligenkänning.

Svenska är huvudspråk i Sverige och ett av Finlands officiella språk, vilket gör svenska till ett pluricentriskt språk. Som Finlands största minoritet har finlandssvenskar rätt till samma tillgång till taligenkänningsverktyg. Det är därför viktigt att utreda hur dessa verktyg behandlar finlandssvenska, som är en regional variant av svenska som talas i Finland. Det är rimligt att anta att taligenkänningsverktyg i Finland som enbart är tränade med sverigevenska eller finska kan göra felaktiga tolkningar och missförstå finlandssvenska på grund av språkliga skillnader.

I min avhandling kommer jag att rikta uppmärksamheten mot pluricentriska språk och undersöka hur dagens taligenkänningsverktyg hanterar finlandssvenska samt hur dessa verktyg kan vidareutvecklas. Utöver det kommer jag även att jämföra finlandssvenska med

sverigesvenska för att få en djupare förståelse för hur bra taligenkänning fungerar. Genom jämförelsen kan jag analysera om verktygen behandlar båda språkvarianterna lika bra och samtidigt undersöka orsakerna till eventuella skillnader som uppstår. Genom att utforska hur verktyg behandlar pluricentriska språk syftar min avhandling att bidra till en bättre förståelse av hur dessa verktyg kan utvecklas för att möta behoven hos finlandssvenska talare. Detta är av betydelse för att stödja tvåspråkigheten och stärka svenskans ställning i Finland.

1.1 Syfte

Syftet med föreliggande studie är att undersöka hur taligenkänningsverktyg hanterar finlandssvenska, som är en icke-dominerande variant av svenska, i jämförelse med sverigesvenska, som är den dominerande varianten. Begreppen behandlas i kapitel 2. Jag kommer även att analysera och identifiera eventuella skillnader mellan de två språkvarianterna. De forskningsfrågor som jag avser att besvara är följande:

- Hur behandlar dagens taligenkänningsverktyg finlandssvenska jämfört med sverigesvenska?
- Vilka orsaker kan ligga bakom eventuella skillnader i tolkningen av finlandssvenska och sverigesvenska?
- Vilka möjligheter finns det för vidareutveckling av taligenkänningsverktyg för att bättre möta behoven hos finlandssvenska talare?

För att utreda hur bra taligenkänningsverktyg tolkar finlandssvenska med avseende på dess regionala variation kommer jag att identifiera de fel som verktyg gör och jämföra resultaten med sverigesvenska. I den andra frågan utforskar jag olika faktorer som kan påverka verktygens tolkning, till exempel uttalsmönster och lexikala skillnader. Jag förväntar mig att finna betydande skillnader i hur taligenkänning hanterar finlandssvenska och sverigesvenska eftersom de två varianterna skiljer sig från varandra särskilt när det gäller talat språk. Genom att kartlägga vilka fel verktyget gör vill jag bidra till diskussionen om hur taligenkänningsverktyg kan vidareutvecklas.

1.2 Material och metod

Materialet i min avhandling är talat språk, dvs. ljud, som jag kan mata i ett taligenkänningsverktyg. För att kunna jämföra finlandssvenska och sverigesvenska med

varandra har jag samlat ljudfiler för båda språkvarianterna. Jag har försökt hitta material där talare med olika regionala bakgrund behandlar ungefär liknande ämnen. På detta sätt kan man försöka minimera olikheter i vokabulären som eventuellt kunde leda till onödiga skillnader vid taligenkänning. Därför är debatter i politik där det talas om aktuella och globala ärenden en bra källa till material. Det är viktigt att det finns åtminstone en grov transkription av talet, så att jag har någonting som jag kan jämföra med det som taligenkänningen producerar. Partiledardebatter och riksdagsplenum finns transkriberade och tillgängliga på nätet så därför kommer jag att använda dem som material i min avhandling.

Då ljudfilerna matas in separat i ett taligenkänningsverktyg får jag verktygets tolkning på både sverigesvenska och finlandssvenska i skriftlig form. Sedan kan jag analysera hur verktyget har tolkat språkvarianterna, finns det t.ex. liknande fel i tolkningarna. Efter att jag har jämfört språkvarianterna med varandra jämför jag även taligenkänningsverktygets produktion med de tillgängliga transkriptionerna för att se om det finns rena misstag i verktygets tolkningar, t.ex. ord som taligenkänningen har tappat bort.

Jag strävar efter att förstå hur taligenkänningsverktyg fungerar och analysera hur användbart material verktyget producerar. Studien kräver såväl kvalitativa som kvantitativa metoder. Kvalitativa metoder används för att analysera resultaten djupare och kvantitativa metoder för att utreda hur autentiskt material taligenkänningsverktyg producerar. Att systematiskt plocka ut och räkna eventuella fel hjälper mig att utvärdera verktyget, medan jämförelser mellan de två varianterna samt mellan tolkningar och transkriptioner hjälper mig att analysera varför eventuella fel och skillnader i taligenkänningsverktygets tolkning uppstår.

1.3 Avhandlingens disposition

Avhandlingen inleds med att jag presenterar bakgrunden och centrala begrepp inom pluricentriska språk samt de viktigaste skillnaderna mellan finlandssvenska och sverigesvenska i kapitel 2. Studier av tal och språkteknologi presenteras i kapitel 3 respektive 4. Den teoretiska bakgrunden kopplas till samhälleliga synpunkter i kapitel 5. Därefter behandlas material och metod närmare i kapitel 6. Resultaten presenteras och analyseras i kapitel 7 och 8 och kapitel 9 avslutar avhandlingen med en sammanfattande diskussion.

2 Pluricentriska språk

Enligt Clyne (1992a:1 f.) introducerades begreppet *pluricentrisk* (*pluricentric*) första gången av Heinz Kloss (1978: 66 f.). Det hänvisar till språk med två eller flera nationella varieteter som skiljer sig från varandra. Skillnaderna mellan de olika nationella varieteterna behöver inte vara stora. Clyne (1992a:2) beskriver skillnaderna så att de tillför smak snarare än innehåll i språkvarianten i fråga. På samma sätt som vid dialekter, kan man tänka, men det som skiljer varieteter av pluricentriska språk från t.ex. regionala dialekter är den höga statusnivån som de har.

Clyne samlade information om världens pluricentriska språk i en antologi 1992. Många pluricentriska språk bildar ett kontinuum med sina nationella varieteter, där språken i fråga används i två eller flera nationer över politiska gränser, vilket är fallet med svenskans varianter, sverigesvenska och finlandssvenska. Det kan dock finnas olika orsaker till att vissa av de pluricentriska varianterna i världen är mer utspridda, t.ex. på grund av emigration, som till exempel varianter av engelska och portugisiska. Pluricentriska språk kan därmed fungera som förenande eller åtskiljande faktor mellan nationer och människogrupper. (Clyne 1992a:1 ff.) En förenande faktor mellan finlandssvenska och sverigesvenska är att svensktalande finländare kan resa till Sverige, tala svenska och bli förstådda. Därtill har vi en del förenande lagstiftning, t.ex. är EU-lagstiftning på svenska gemensam för Sverige och Finland (Reuter 2015b:22). Gemensam lagstiftning kan dock också ses som en åtskiljande faktor. Terminologi i lagtexter följer många gånger terminologi som är etablerad i Sverige, inte i Finland, vilket kan leda till tankar om att finlandssvenska är en mindre värdefull språkvariant.

Eftersom pluricentriska språk innebär flera språkvarianter respektive nationer, uppstår också frågan om vem som har den högsta maktpositionen. Clyne (1992b:455) påpekar att pluricentricitet nästan alltid leder till obalans, vilket betyder att någon eller några nationer har en högre maktnivå än de andra. Vem som får den högsta maktpositionen kan avgöras på grundval av vem som talar den äldsta eller ursprungliga språkvarianten, vem som har en större nation eller största delen av talarna eller vem som har uppnått en maktposition tack vare ekonomisk eller politisk styrka. För att kunna beskriva sådana här maktförhållanden har Clyne (1992b:459) lanserat två begrepp: dominerande (*dominant*) variant och ”övriga” varianter, som senare kommit att kallas dominerande och icke-dominerande (*non-dominant*) varianter för enhetlighetens skull (WGNDV, 2012). Maktpositioner kan också förändras (Clyne

1992a:3), dvs. en dominerande variant av ett pluricentriskt språk kan bli en icke-dominerande variant av språket i fråga och tvärtom.

Som speciellt intressanta frågor relaterade till pluricentriska språk listar Clyne (1992a:4) bland annat att identifiera vilken status olika språkvarianter har och om någon variant dominerar, och varför är det just den varianten i så fall. Nätverket *International Working Group on Pluricentric Languages and their Non-dominant Varieties* (WGNDV) arbetar med att identifiera och dokumentera alla de icke-dominerande varianterna av världens pluricentriska språk. Just nu är det kring 50 språk i världen som kan räknas som pluricentriska, vilket resulterar i ungefär 290 nationella varieteter. Kartläggningen pågår och vem som helst kan bidra genom att lämna in information om någon av de 290 nationella varieteterna på WGNDV:s webbplats. (WGNDV 2023)

2.1 Svenska som ett pluricentriskt språk

Det finns de som argumenterar att finlandssvenska skulle kunna ses som ett eget språk, eftersom vi vet att det finns olikheter mellan svenskan i Sverige och svenskan i Finland. Om detta skulle vara fallet kunde vi i Finland sluta följa sverigesvenskans normer och skapa våra egna finlandssvenska normer (Reuter 2015b:36). Men Reuter (1992:111 f.) argumenterar för att de gemensamma dragen för finlandssvenska och sverigesvenska väger tyngre. I stort sett har varianterna gemensam ortografi, vokabulär, morfologi och kultur, och därför skulle det inte vara realistiskt eller ens önskvärt att se finlandssvenska som ett eget språk (Reuter 2015b:36). Engstrand (2012:133) är inne på liknande spår och menar att två språkvarianter som delar skriftspråk, litteratur och till och med massmedier kan räknas som samma språk, även om de används på olika sidor om en nationsgräns. Därför behandlar jag sverigesvenska och finlandssvenska som två varianter av ett och samma språk i min avhandling.

Innan dagens Sverige fick sina nuvarande gränser hörde delar av Sverige tidvis till Danmark eller Norge, vilket fortfarande kan höras i dialekter som talas nära statsgränsen till Danmark respektive Norge. Det handlar om dialekter som har sitt ursprung i urnordiskan som talades i hela Norden innan de nuvarande statsgränserna bildades (Engstrand 2012:107, 112, 126). Det finns därmed skillnader i talat språk och flera dialekter som skiljer sig från standardspråket. För att kunna bli en enhetlig nation var det nödvändigt att också utveckla ett svenskt skriftspråk (Engstrand 2012:107 f.). Religiösa texter behövde vara tillgängliga på svenska under 1500-talet och därmed blev översättningen av Nya testamentet en viktig milstolpe.

Senare, år 1842, infördes folkskolan i Sverige och då var det standardsvenska alla svenska barn skulle läsa och skriva på (ibid.) Skriftspråket präglas alltså inte av skillnader i talat språk utan är standardiserat i hela språkområdet.

Finland har varit under svenskt och ryskt styre. Fram tills senare hälften av 1800-talet förblev det svenska standardspråket som utbildnings- och administrationsspråk i Finland. Detta betyder att svenskan var det dominerande språket även då Finland var en del av Ryssland. På 1900-talet började finskan få starkare fotfäste i Finland, i synnerhet efter självständigheten 1917. År 1922 deklarerades både finskan och svenskan vara Finlands likvärdiga officiella språk. Även om finskan sedan dess har haft en högre status i Finland än svenskan, har det alltid funnit en svenskspråkig minoritet i Finland. Tvåspråkigheten i Finland syns bland annat i att lagstiftning skrivs parallellt på finska och svenska och utbildning ges på svenska på alla nivåer än i dag. Den historiska bakgrunden samt närheten till Sverige förklarar hur svenskan har lyckats att behålla sin position i Finland. (Reuter 1992:101–103)

Enligt Sveriges språklag har landet endast ett huvudspråk, svenska, medan det i Finland finns två officiella språk, finska och svenska (Reuter 1992:101). Svenska, med sina två varianter, är därmed ett av de ca 50 språken som finns med på listan över världens pluricentriska språk. Eftersom de flest svensktalande är bosätta i Sverige har svenskan i Sverige klart en högre maktposition i förhållande till svenskan i Finland (Clyne 1992b:459 f.).

Reuter (1992:105, även Reuter 2015a:20) påpekar att många svenskar inte alls är medvetna om hur mycket gemensam historia Sverige och Finland har och att något som heter *finlandssvenska* över huvud taget existerar. Alternativt tror svenskar att finlandssvenskar är finskspråkiga finländare som har flyttat till Sverige. Clyne (1992b:459 f.) konstaterar att detta är ett till tecken på att det är Sverige som har den högre maktpositionen, eftersom det är vanligt att de som talar den dominerande språkvarianten inte är så insatta i de icke-dominerande språkvarianterna.

Trots olika maktförhållanden, eller kanske på grund av dem, är det viktigt att visa erkännande även för de icke-dominerande språkvarianterna. Clyne (1992b:455) argumenterar för att det annars kan hända att den icke-dominerande språkvarianten så småningom utvecklas bort från den dominerande varianten, t.o.m. till den mån att varianterna inte längre kan identifieras med varandra. Med tanke på detta skulle det till och med vara kontraproduktivt om finlandssvenskan skulle utvecklas till ett eget språk som inte längre skulle kunna användas för

att kommunicera på svenska med i Sverige (Reuter 2015b:12). Därför är det viktigt att vi i Finland t.ex. kontrollerar användningen av finlandismer och säkerställer att det svenska skriftspråket i Finland utvecklas i samma riktning som i Sverige (Reuter 1992:113 f. och Reuter 2015b:12). Reuter (2015b:32) betonar att det således inte räcker med att vi försöker undvika vissa finlandismer, utan utvecklingen av skriftspråket i Finland bör aktivt följa utvecklingen i Sverige. Finlandssvenska får alltså ha sina egna särdrag, men särskilt skriftspråket bör följa sverigesvenskans regler och riktlinjer (Reuter 2015b:30), vilket återigen visar sverigesvenskans maktposition (Clyne 1992b:459 f.).

Även om vi strävar efter att utveckla det svenska skriftspråket i Finland i samma takt som i Sverige förekommer det som Reuter (1992:106) påpekar fler skillnader än vi kanske tror i skriven finlandssvenska jämfört med skriven sverigesvenska. Skillnader finns särskilt i ordförrådet (Reuter:ibid.), vilket framgår ännu tydligare i tal. Reuter (1992:104) konstaterar att vokabulär, dvs. ordval, och idiomatiska uttryck lätt avslöjar en finlandssvensk. Trots att vi bör undvika onödiga finlandismer finns det en del finlandismer som är nödvändiga för kommunikationen (Ivars 1991:5) eftersom det samhälleliga sammanhanget är ett annat (Reuter 1992:111 och Reuter 2015b:12, 23). Det är fortfarande möjligt att hitta skrivna svenska texter där man inte kan vara säker om skribenten är sverigesvensk eller finlandssvensk, men hör vi talad svenska är det mycket lättare att gissa rätt. (Reuter 1992:passim., Reuter 2015b:12 och Engstrand 2012:105 f.) Därför är det utan vidare mer spännande att studera talat språk när det kommer till pluricentriska språk (Engstrand 2012:105 och Reuter 2015b:12). I dag läggs det allt mer fokus på språkbruk i sociala sammanhang samt likheter och olikheter mellan nationella varieteter av pluricentriska språk. I denna avhandling fokuserar jag på svenska som pluricentriskt språk utifrån uttalsskillnader, som behandlas i nästa kapitel.

2.1.1 Uttalsskillnader mellan finlandssvenska och sverigesvenska

Vi reagerar inte bara på vad folk säger utan också snabbt hur folk säger någonting. En viss sorts brytning kan avslöja utlänningar ganska fort, det kan räcka med en sekund, dvs. ett enda ord (Engstrand 2016:7 f.). Ibland kan vi till och med identifiera från vilket land en person kommer från eller, baserat på dialekten, om en person kommer t.ex. från norr eller söder. Så här enkelt är det oftast också att besluta om en svensktalande är hemma från Sverige eller Finland (Reuter 2015b:20).

Nu ska vi fokusera på skillnader mellan finlandssvenskt och sverigesvenskt uttal utgående från en översikt som Reuter (2015a:19–34) har sammanställt. Reuter påpekar att översikten är användbar just för att jämföra fonologiska och morfologiska skillnader mellan de två varianterna. Med finlandssvenskt uttal här avses överregionalt finlandssvenskt talspråk, dvs, ett skriftspråksnära tal, och med sverigesvenskt uttal ett centralsvenskt standarduttal.

Finlandssvenskan och sverigesvenskan har precis samma vokalfonem. Trots detta manifesteras olikheterna på flera punkter på olika sätt, såsom i de enskilda vokalernas kvalitet och hur vissa vokaler förekommer i enstaka ord. Ett exempel på hur vokalkvaliteten skiljer sig mellan språkvarianterna är att u-ljudet i finlandssvenskan är lika slutet och rundat som kort och långt uttal, dvs. vokalen bildas alltid på samma ställe i munnen. I sverigesvenskan uttalas däremot ett kort u-ljud längre bak i munnen än ett långt u-ljud. Detta betyder att orden *ful* och *full* i sverigesvenskan får olika vokalkvalitet men i finlandssvenskan är vokalkvaliteten densamma. Ett annat exempel på skillnader i vokalkvaliteten är att de slutna vokalerna *i*, *y*, *u* och *o* i sverigesvenskan oftast får en konsonantisk slutfas när de uttalas som långa (såsom /ij/, /yj/ osv.), men inte i finlandssvenskan. (Reuter 2015a:21–24)

Vokalerna i enskilda ord kan illustreras med vokalerna /o/ och /å/. Det finns en hel del ord som i finlandssvenskan uttalas med långt /o:/ och i sverigesvenskan med långt /å:/, t.ex. *mikrofon*, *telefon*, *tropisk* och *anekdot*. Fenomenet förekommer också omvänt, ord som *bonus* och *fokus* uttalas i finlandssvenskan med vokalen /å/ och i sverigesvenskan med /o/. Sedan finns det en mängd ord som i sverigesvenskan uttalas med kort /o/ och i finlandssvenskan med kort /å/, som *boxas* och *morsk*. Att växla mellan dessa vokalljud är inte ett stort problem eftersom ordens betydelse inte ändras. Allmänt taget kan det dock konstateras att å-uttal är vanligare i finlandssvenskan respektive o-uttal i sverigesvenskan. (ibid.)

De flesta ord med /eu/ i finlandssvenskan uttalas som diftongen /eu/ (t.ex. *euro* och *eutanasi*), men det gamla uttalet /öj/ lever fortfarande kvar i uttalet av *farmaceut* och *terapeut*. I sverigesvenskan har man kvar det gamla /öj/ i namn som *Reuter*, men i de flesta fall uttalas /eu/ som /ev/, t.ex. /evro/ och /terapevt/. Både sverigesvenskans och finlandsvenskans vokalsystem avviker från finskans vokalsystem vid t.ex. den slutna mellanvokalen [ʉ] som inte har någon motsvarighet i finskan. Om vi tänker på fonologisk kvalitet så uttalas vokaler över huvud taget spändare i sverigesvenska än i finskan, som påverkar finlandssvenskan, särskilt i Helsingforsområdet. (Reuter 2015a:21–24)

Skillnaderna i konsonanter beror på inverkan från finska, och därtill finlandssvenskans ålderdomliga drag. Finlandsvenskans tonlösa klusiler *p*, *t* och *k* är oaspirerade, dvs. de saknar det h-liknande ljud som är typiskt för sverigesvenska och många andra germanska språk. Detta kan leda till att dessa klusiler för en sverigesvensk kan låta som de tonande motsvarigheterna *b*, *d* och *g*. Supradentaler förekommer i vissa finlandssvenska dialekter, men allmänt kan man säga att ord som *kort*, *bord* och *barn* i finlandssvenskan uttalas med hörbart *r*. Däremot uttalas *rn*, *rt*, *rd*, *rs* och *rl* i sverigesvenskan med supradentala konsonanter [ɳ], [ʈ], [ɖ], [ʂ] respektive [ɭ], så att de två konsonanterna smälter samman. (Reuter 2015a:24 ff.)

Finlandssvenskans sje-ljud är en tonlös post-alveolär frikativa [ʃ], medan sverigesvenskans [ɧ] är rundare och bildas med tungryggen mot den mjuka gommen, alltså längre in i munnen. Detta leder till att finlandssvenskans sje-ljud för sverigesvenskar låter mer som deras tje-ljud. Sverigesvenskans tje-ljud är, på samma sätt som sje-ljudet, en tonlös frikativa. Artikulationssättet är således detsamma för sverigesvenskans sje- och tje-ljud men artikulationsställningen är en annan. Sverigesvenskans tje-ljud bildas med tungbladet mot den bakersta delen av tandvallen och tungryggen mot den hårda gommen [ɕ]. Finlandssvenskans tje-ljud är däremot en affrikata, vanligtvis med tydligare t-ljud [t͡ʃ]. Ord som *kiosk*, *kilo* och *arkitekt* uttalas med tje-ljud i sverigesvenskan men i finlandssvenska oftast med *k*. (ibid.)

Andra skillnader vid konsonanter är att s-ljudet i finlandssvenskan ofta uttalas mindre spetsigt än i sverigesvenskan och att förbindelsen *ng* i sverigesvenskan uttalas som [ŋg], alltså med hörbart *g*, då det i finlandssvenskan oftast reduceras till bara [ŋ], t.ex. i *bingo*, *fungera* och *lingvist*. (Reuter 2015a:24 ff.)

Kortstavighet, som också är ett ålderdomligt drag, är mycket vanligt i finlandssvenskan. Det visar hur även kvantiteten skiljer sig mellan de två språkvarianterna. Kortstavighet betyder att en kort betonad vokal inte följs av någon konsonant alls eller av en kort konsonant i öppen stavelse, t.ex. i ord som *nu*, *mina* och *föredrag* kan få ett kortstavigt uttal. Kortstavighet förekommer mer i informella sammanhang och kan ses som ovårdat språk, men total avsaknad av kortstavighet låter inte heller som idiomatiskt finlandssvenska. Sverigesvenskan har däremot en kvantitetsregel som specificerar att varje betonad stavelse endast kan ha ett långt ljud och att långa ljud endast kan förekomma i betonade stavelser. Därtill finns det ord som i finlandssvenskan vanligen har ett uttal med lång vokal och kort konsonant där sverigesvenskan tvärtom har kort vokal och lång konsonant, som *honung*, *ensam*, *kedja* och *domare*. (Reuter 2015a:26 ff.)

Intonation är vidare en av de tydligaste skillnaderna som snabbt framgår av talet (Reuter 1992:104). I finlandssvenskan finns det bara en accenttyp, monoton satsintonation, i likhet med finskan. Däremot finns det två typer av ordaccent i sverigesvenskan, akut och grav accent. Dessa accenter är en orsak till att sverigesvenskan ofta låter musikaliskt. De två accenttyperna särskiljer också betydelser, t.ex. *tomten* (av *tomt*) har en tontopp, dvs. akut accent, medan *tomten* (av *tomte*) har två tontoppar, dvs. grav accent. (Reuter 2015a:28 f.)

Betoningen ser i stort sett lika ut i finlandssvenska och sverigesvenska, men det finns några viktiga skillnader. I sammansättningar som *förhandsinställning* och *universitetsbokhandel* uttalas det senare ledet (*inställning*, *bokhandel*) med grav accent i sverigesvenska. Huvudbetoningen i dessa sammansättningar blir således på det första ledet och bibetoningen på den sista stavelsen i sverigesvenskan; *'förhandsin,ställning*, *universi'tetsbok,handel*. I finlandsvenskan behåller varje led sin betoning även i sammansatta ord; *'förhands,inställning*, *universi'tetsbokhandel*. Följden blir att ett ord som *statsrådslag* i finlandssvenskan kan få två olika betoningar beroende på om man talar om en lag om statsråd (*'statsråds,lag*) eller om ett rådslag mellan stater (*'stats,rådslag*). Detta kan däremot inte ske i sverigesvenskan, vilket visar hur betoningen i sammansatta ord är starkt relaterad till ordaccenter. (Reuter 2015a:29 ff.)

Orden *alibi*, *budget* och *ursäkt* är exempel på ord som i sverigesvenskan betonas på första stavelsen men på senare stavelsen i finlandssvenskan. Vissa ord som slutar på *-iv* har alternativt betoning på första eller sista stavelsen och i de fallen verkar finlandssvenskan oftare än sverigesvenskan flytta betoningen till första stavelsen, t.ex. i ord som *definitiv*, *effektiv* och *innovativ*. Ord som *administrativ*, *meditativ* och *alternativ* betonas ofta på slutstavelsen i sverigesvenskan men får betoning på första stavelsen i finlandssvenskan. Detta är ytterligare ett exempel på finlandssvenskans kortstaviga uttal som nämnts ovan. (ibid.)

Utelämnande av preteritumändelsen *-de* (som i *kastade*), supinumändelsen *-t* (som i *kastat*) och den bestämda neutrumändelsen *-t* (som i *huset*) har tidigare varit vanligt i både centralsvenskt och finlandssvenskt talspråk. Numera har ändelserna dock börjat komma tillbaka i sverigesvenskan och hörs i vardagligt talspråk; i finlandssvenskan hörs de vanligen inte. På liknande sätt är det vanligare i finlandssvenskan att säga *int'*, *måst' sku'* och *satt'* i stället för *inte*, *måste*, *skulle*, och *satte*. Orden *är* och *och* förkortas också gärna i finlandssvenskan, men i sverigesvenskan hör man ganska ofta att dessa småord uttalas med hörbara slutkonsonanter. (Reuter 2015a:31)

Det finns alltså två typer av skillnader. En del är fasta, t.ex. finlandssvenskans avsaknad av grav accent och avvikande vokalkvalitet för a- och u-ljud, men en del skillnader kan talare välja mellan, t.ex. kortstavigt uttal och om man betonar ord enligt sverigesvenskans eller finlandssvenskans mönster. Således har finlandssvenskar alltid möjligheten att variera sitt tal och uttal, t.ex. om de talar med svenskar. Att städa bort finlandismer gör man kanske medvetet, men man kan också anpassa sitt tal omedvetet, t.ex. genom att plötsligt betona ord enligt sverigesvenskan. (Reuter 2015a:31 ff.)

Tack vare tidigare studier vet vi att de största och tydligaste skillnaderna mellan finlandssvenska och sverigesvenska hörs i uttalet. Även om det finns många och olika slags skillnader i sverigesvenskt och finlandssvenskt uttal betyder det inte nödvändigtvis svårigheter i kommunikationen. Enligt Reuter (2015a: 20, 32 f.) upplevs vårdad finlandssvenska som en relativt lättförståelig variant i Sverige. Det är inte heller så att finlandssvenskar strävar efter att låta som sverigesvenskar. Finlandssvenskar som bor och lever i Finland har inte så stor nytta av att lära sig grav och akut accent, särskilt eftersom finlandssvenskan anses vara lätt att förstå (Reuter 1992:105). Endast några av de viktigaste fonetiska reglerna i sverigesvenskan kan höras i formellt finlandssvenskt tal. Ett exempel på detta är assimilation mellan hård och mjuk *k* i ord som *kort* och *köra*.

Detta kapitel har behandlat skillnader mellan finlandssvenska och sverigesvenska i standardiserat, vårdat uttal. Vi bör ändå komma ihåg att båda språkvarianterna varierar i viss mån. Därtill måste vi ta hänsyn till att vissa uttalsdrag som ofta räknas enbart som sverigesvenska förekommer i en del av Svenskfinland, t.ex. supradentaler på Åland. Engstrand (2012:105) påpekar att det inte heller handlar bara om dialekter och andra regionala skillnader utan språkliga skillnader förekommer också på ålder, kön, utbildning och samhällsställning.

Uttalsskillnaderna får alltså talare att låta olika men påverkar normalt sett inte i förståelsen mellan finlandssvenska och sverigesvenska. Skillnader i uttal är en av orsakerna som gör pluricentriska språk intressanta att studera, och särskilt vad skillnaderna betyder med tanke på taligenkänning. I och med att det finns rikligt med skillnader kan vi med stor sannolikhet anta att finlandssvenska och sverigesvenska orsakar vissa svårigheter vid taligenkänning. Därför är det framför allt sådana här utmaningar som jag antar mig finna i mitt material.

Betoning, intonation och ordaccenter räknas till det som kallas prosodiska drag. Engstrand (2004:173) definierar prosodiska drag som egenskaper som bildar melodi och rytm i vårt tal, alltså egenskaper som överlagrar på vokaler och konsonanter. Engstrand (ibid.) konstaterar vidare att sverigesvenskan har rikligt med prosodiska drag jämfört med världens andra språk, vilket inte är fallet med finlandssvenskan som ligger närmare finskans fonologi (Kuronen och Leinonen 2001:125 och Reuter 1992:106). Detta framgår även tydligt i Reuters (2015a:19–34) sammanställning som presenterades ovan. Kuronen och Leinonen (2001:126) håller med om att det enklaste sättet att identifiera svensktalare som finlandssvensk eller sverigesvensk är just med hjälp av prosodiska drag. Talet har således en prosodisk struktur i sig. De två följande kapitel fokuserar på de teoretiska utgångspunkterna för studien. Fonetik diskuteras i kapitel 3 och språkteknologi i kapitel 4.

3 Fonetik

Bakgrunden till min avhandling är delvis i fonetik, delvis i språkteknologi. Att kombinera dessa två fält hjälper mig att förstå hur taligenkänning fungerar samt hur vi kan förbättra taligenkänningsverktyg. Nedan ska jag gå genom de genom viktigaste utgångspunkterna för min avhandling och ger en översikt över tidigare forskning. Detta kapitel fokuserar på fonetik och kapitel 4 på språkteknologi.

Fonetik är ett tvärvetenskapligt fält som studerar språk och människans möjlighet att producera språkljud samt t.ex. hur vokaler och konsonanter formas. Fonologi däremot fokuserar på hur vi särskiljer betydelser med hjälp av skillnader i språkljud. Det är viktigt att notera att fonetik inte beskriver något specifikt språk, till skillnad från fonologi som är språkspecifikt. I det här kapitlet beskriver jag hur fonetik hänger samman med talteknologi och hur genom att förstå fonetiska skillnader i språk kan utveckla taligenkänning.

För att bättre kunna förstå talteknologi bör vi först förstå vad tal är. Tal skapas när luft passerar genom vår talapparat, och talapparaten förändras enligt de ljud som vi vill producera och ljudsignalerna bildar ord. Engstrand (2016:18 f.) illustrerar vad tal är i förhållande till skrift genom ett enkelt experiment. I experimentet spelade Engstrand först in enskilda vokaler och konsonanter och satt sedan ihop de korta inspelningarna så att de bildade frasen *vilka härliga pannkakor*. Det är uppenbart att en fras som denna inte kan låta så hemskt naturlig. Tal är inte bara enstaka konsonanter och vokaler efter varandra. Om vi ser på naturligt producerat tal ser vi att det handlar om ett kontinuum där det inte finns några tydliga gränser mellan språkljuden, såsom det finns mellanslag i text. I naturligt tal glider språkljuden från ett till annat, samtidigt som de påverkar varandra. Dessutom förekommer det med mycket mer, t.ex. suckar, smackar och andra biljud i naturligt tal. (Engstrand 2016:18 f.) Sedan är att samtala spontant mycket annorlunda än att till exempel hålla tal. Engstrand (2012:9) påpekar att meningar i spontant tal blir oavslutade. Vi pratar i munnen varandra och ordföljden och grammatiken blir lite hur som helst.

Talspråklig kommunikation bildas av två huvudaktörer, den som talar som skickar en akustisk talsignal och den som lyssnar och tolkar den akustiska signalen (Engstrand 2004:41). Kommunikationen bygger därför på förståelse av den akustiska signalen ur både talarens och lyssnarens perspektiv. Då vi talar ett gemensamt språk har vi oftast inga problem att förstå varandra tack vare människans utvecklade språkförmåga. Att förstå tal, dvs. att plocka alla

relevanta ljud och bilda ord, går så fort och automatiskt att vi sällan behöver tänka på det (Engstrand 2016:28). Biljud orsakas bland annat av att vår talapparat är bildad också för andra ändamål. Vi tuggar, sväljer och andas med samma muskler som vi producerar tal med. Tittar vi på detta ur ett biologiskt perspektiv så är att tala egentligen en sekundär funktion (Engstrand 2004:75). Engstrand (2016:22, 67) påpekar att den mänskliga hjärnan lätt kan filtrera bort ljud som inte är relevanta för förståelsen, så som suckar, trots att vi hör dem.

Å ena sidan kan vi plocka bort biljud som inte är relevanta för förståelsen men å andra sidan hör vi också mycket sådant som vi inte säger. Om vi har ont om tid kan vi behöva förkorta ord. I stället för att uttala varje språkljud i t.ex. ordet *naturligtvis* kan vi säga något i stil med *natulitvis*, *natultvis*, *natutvis* eller något ännu kortare (Engstrand 2012:13). Att förkorta och förenkla ord och hela meningar på detta sätt förekommer hela tiden, det är helt enkelt något vi alla gör även om vi skulle ha all tid i världen (Engstrand 2016:63 och 2012:12, 16.). Engstrand (2016:27 f.) ger exempel ur radion där det i dag är vanligare att tala mer avslappnat, vilket kan betyda att frasen *jag har faktiskt skrivit* kan få ett uttal som låter ungefär som *jaafak'srivit*. Där har ordet *faktiskt* förkortats till bara några få bokstäver, *fak(tiskt)*. Ett sätt att förenkla uttalet är assimilation, alltså att låta språkljuden smälta samman och anpassa sig efter omgivningen. I praktiken betyder detta att vi i stället för att säga *en bok* ofta säger *em bok* eller till och med *mbok*, där *m* är en anpassning till *b*. (Engstrand 2016:15) Konsonanterna *b* och *m* bildas bilabialt, alltså med båda läpparna, och därför är det logiskt att assimilera språkljuden i stället för att uttala *n* som bildas på ett helt annat sätt. Att hoppa mellan alla möjliga artikulationssätt och -ställningar är helt enkelt för tidskrävande.

Även om vi håller på med att förkorta och sammandra ord menar Engstrand (2016:22 och 2012:12 f.) att det går lätt att sätta ihop allt till ord och meningar. Våra hjärnor gör det så automatiskt att vi endast behöver höra en början så fyller hjärnan i resten (Engstrand 2016:21 ff.). Engstrand (ibid.) menar att det egentligen skulle vara svårare att inte hoppa över språkljud och alltid genomartikulera varje ord. Typiskt för talets förenkling är t.ex. att börja hoppa över de obetonade stavelserna först (Engstrand 2012:23 f.). Engstrand (2016:28) påpekar vidare att det att förkorta, förenkla och dra ihop ord är inte samma sak som sludder.

Engstrand (2004:14 f.) sammanfattar varför det kan vara en bra idé att ta med fonetiker i utvecklingen av språk- och talteknologi: en del av det fonetiker gör är att studera vad i uttal som utgör de avvikande drag som t.ex. skiljer dialekter från standardspråken. Fonologi bidrar därmed till att förstå vad och vilka skillnader det är som maskiner ska lära sig. Enligt

Engstrand (ibid.) är fonologi kombinerad med talteknologi därför ofta en mycket framgångsrik kombination. Detta stöds även av det Forsberg (2003:7) och Ivars (1991:5) konstaterar. Men hur ska vi lära maskiner tolka vilka ljud i talet som är relevanta och vilka som inte är det? Det är bland annat sådana här frågor man försöker lösa då man jobbar med talteknologi och taligenkänning.

De första systemen som liknar dagens taligenkänning utvecklades 1952 av *Bell Laboratories* och baserades på akustisk fonetik. Idén var att förklara hur de grundläggande språkljuden i ett språk manifesteras i naturligt tal och vilken akustik språkljuden har i mänskligt tal. Talets och språkljudens akustik påverkas bland annat av språkljudens artikulationsställning och -sätt. Akustisk fonetik användes som grund för att utveckla ett system för att identifiera siffror. Frekvensområdet för varje språkljud kodades med en siffra och således, när en informant talade, kunde systemet identifiera språkljudens frekvenser och kopplade därefter frekvenserna till siffror. (Juang och Rabiner 2005:6 och Sonix 2024) Detta visar klart och tydligt hur man med hjälp av att förstå fonologi kan göra viktiga framsteg inom talteknologin. Språkteknologi, taligenkänning och fältets framsteg samt genombrott behandlas närmare i nästa kapitel.

4 Språkteknologi

Språkteknologi är en hyperonym som täcker all interaktion mellan datorer och språk, bl. a. textbehandling, översättning och taligenkänning. Språkteknologi är alltså teknologi som utnyttjar det mänskliga språket. Nedan följer en koncis genomgång av hur språkteknologi som fält har kommit till samt hur det har utvecklats genom tiderna. Efter det behandlas taligenkänning och dess utmaningar närmare.

Språkteknologi och taligenkänning är begrepp som har blivit alltmer bekanta för fler människor under det senaste decenniet. Juang och Rabiner (2005:1–4) sammanfattar att tanken från början har varit att talteknologin skulle hjälpa oss att arbeta på ett snabbare och effektivare sätt. Denna tanke lever kvar ännu i dag. Som mål har forskare haft att vi ska kunna styra maskiner med röstkommandon och lita på att maskinerna faktiskt arbetar enligt givna instruktioner. Det kan konstateras att språkteknologi är ett relativt nytt fält. Dock bör vi komma ihåg att mycket i språkteknologi har sitt ursprung i språkvetenskapen, som är ett betydligt äldre vetenskapsområde. En del i språkteknologin bygger på datateknik, vilket på liknande sätt som språkteknologi är ett yngre fält. Trots detta har språkteknologi snabbt tagit stora steg.

Redan för fem decennier sedan fanns det intresse att utveckla en maskin som producerar tal, till skillnad från att igenkänna tal. Från populärkulturen lyfter Juang och Rabiner (2005:1–4) fram *Star Wars* R2-D2 som ett exempel på en sådan maskin. Först lite senare flyttades fokus till att utveckla maskiner som inte nödvändigtvis behövde tala utan förstå och tolka mänskligt tal. På 1960-talet utvecklades det flera japanska system som fokuserade på att igenkänna tal (Juang och Rabiner 2005:8 f.). Bland dessa japanska uppfinningar fanns en vokal- och fonemigenkänning samt den första talsegmenteraren som var konstruerad för att dela upp talsignalen i ljudsegment så att talade ljud kunde bearbetas (Sonix 2024). Lite senare i mitten av 1970-talet utvecklade amerikanen Tom Martin något som kan tänkas vara den första produkten som automatiskt igenkände mänskligt tal. Produkten hette *VIP-100 systems* och transporttjänsten *FedEx* använde den för att sortera paket på ett transportband. (Juang och Rabiner 2005:8 f.)

Utvecklingen slutar inte där, utan under 1990- och 2000-talet har teknologin gått framåt och utvecklats så att maskiner och datorer förstår det mänskliga talet ännu bättre. Det har skapats stora databaser med mycket stora mängder ord, dvs. ordböcker och stora språkmodeller, som

datorer använder för att kunna tolka det mänskliga talet. Detta har bland annat revolutionerat översättningen och gett upphov till många verktyg för översättare, till exempel översättningsprogram och -minne. Därmed har det kommit ny teknik och olika förbättringar som gör det lättare för datorer att tolka och förstå tal. (Juang och Rabiner 2005:20) Detta har även påverkat hur fonetiker jobbar (Engstrand 2004:41). Specialkonstruerade maskiner och olika datorprogram har utvecklats för att bättre kunna förstå, spela in och analysera akustiska ljud och talsignaler. I denna avhandling kommer jag dock inte behandla dessa aspekter desto närmare utan jag fokuserar på taligenkänning.

4.1 Taligenkänning

Taligenkänning är ett område inom språkteknologi som gör det möjligt att omvandla talat språk till text, alltså till en form som maskiner förstår (Jurafsky och James 2023). Därför fokuserar taligenkänning på att bearbeta talat språk. Taligenkänning, även förkortad ASR (*automatic speech recognition*), innebär helt enkelt att man omvandlar ljudspår till text. Det är dock viktigt att förstå att tal-till-text-tekniken inte är exakt samma sak som ASR, utan ASR sträcker sig bortom enkel transkription och kräver maskininlärning (Sonix 2024), som jag kommer att förklara närmare i kapitel 4.1.1. I detta delkapitel försöker jag klargöra hur fonetik och språkteknologi konkret kombineras i taligenkänning samt diskuterar de viktigaste problemen inom området.

Det finns flera orsaker till att vi över huvud taget vill dra nytta av och utveckla taligenkänning. Som Juang och Rabiner (2005:1–4) konstaterar ovan är ett av dem den hjälp och effektivitet som teknologin ger; kommunikation med röstkommandon frigör händerna för andra saker. Detta stöds också av Jurafsky och James (2023) samt Forsberg (2003:2). En annan central orsak är att det alltid har varit viktigare att uttrycka sig muntligt än att uttrycka sig skriftligt. Engstrand (2012:20 f.) påpekar att barn lär sig tala tidigare och snabbare än läsa och skriva. Att kommunicera via talat språk är alltså naturligt för oss människor. Arora och Singh (2012:34) tillägger att tala också är mycket snabbare jämfört med att skriftligt kommunicera via ett tangentbord.

Forsberg (2003:7) ger en koncis introduktion till taligenkänning genom att illustrera varför taligenkänning är svårt. Aspekter som han listar är bland annat kroppsspråk, hur talat språk skiljer sig från skrivet språk och mängden variation som beror på talarnas ålder, kön, anatomi, talhastighet samt regionala och sociala dialekter. Arora och Singh (2012:37) nämner därtill

bakgrundsljud, taluppfattning och kanalvariabilitet, dvs. kontext där den akustiska signalen yttras såsom mikrofoner eller vad som helst som påverkar innehållet i den akustiska signalen. Ett av de centrala problemen inom taligenkänning är därmed att tal egentligen är mycket mer än språkljud och ord (Arora och Singh 2012:36). Vi kan t.ex. använda prosodiska drag för att betona att det vi säger är viktigt (Forsberg 2003:3) eller använda kroppsspråk för att signalera om vi är på gott humör eller inte.

Med andra ord är en av de största utmaningarna språkets variation. På detta fick vi även se exempel i kapitel 3 där jag beskrev hur förkortningar, förenklingar samt assimilation varierar inom talspråket. Detta innebär att talets akustik aldrig är exakt densamma. Om vi har en mängd informanter som uttalar samma ord kan vi enligt Juang och Rabiner (2005:20) finna betydande skillnader i talets akustik, beroende på de orsaker som listats ovan. O'Shaughnessy (2008:2966) klargör vad detta betyder i praktiken. Låt oss tänka att vi har en ASR-modell och tränar modellen med en person som upprepar ett och samma ord gång på gång i ett tyst utrymme där det inte finns bakgrungsbuller. Sedan vill vi testa hur bra vår modell fungerar genom att testa den med samma person och i samma utrymme som vi använde för att träna modellen. Denna gång upprepar personen olika test-ords böjningsformer, alltså inte exakt samma ord som användes i träningen. Här kan vi ändå förvänta oss utmärkta resultat, eftersom vår ASR-modell troligtvis kan tolka de olika böjningsformerna korrekt. Däremot om vi testar samma ASR-modell med en annan person, en annan mikrofon och har med bakgrundsbuller, märker vi direkt att modellens prestanda försämras (O'Shaughnessy 2008:2966).

Hur kan taligenkänning över huvud taget fungera då det finns så mycket variation i talat språk? Juang och Rabiner (2005:20) presenterar en akustisk modell som färdigställdes i mitten av 1980-talet. Modellen blev en central metod för taligenkänning och är det fortfarande i dag. Modellen heter HMM (*Hidden Markov Model*) och fungerar på basis av ett sannolikhetsmått. Då det finns mycket variation i talat språk och dess akustik, kan sannolikhetsmåttet ta hänsyn till variationen. Sannolikhetsmåttet använder HMM:s språkliga modell (dvs. ordböcker och stora språkmodeller) för att kunna framställa den mest sannolika tolkningen av talet i skriftlig form. Ordböcker och stora språkmodeller bidrar framför allt till att koppla struktur i språk till artikulation och uttal. Att kunna införliva HMM i ordböcker och språkmodeller var ett mycket viktigt teknologiskt genombrott. (Juang och Rabiner 2005:20) Under 1980-talet kunde transkriptionssystemet *Tangora*, som använde HMM, känna igen 20 000 ord på engelska (Sonix 2024). Detta illustrerar hur taligenkänningsverktyg redan då

kunde igenkänna ett ganska omfattande ordförråd och hur HMM drev framåt utvecklingen av automatisk taligenkänning under 1980-talet.

På 1990-talet började statistik analys ta över utvecklingen och den första kommersiella programvaran för taligenkänning lanserades. År 2007 lanserade *Google* sin egen teknik för röstigenkänning, vilket gjorde tekniken tillgänglig för så gott som alla. Under 2010-talet tog taligenkänning så stora tekniska språng, bland annat genom djupinlärning, att det blev allmänt vedertaget. (Sonix 2024) Djupinlärning och andra aspekter i taligenkänning diskuteras kortfattat i nästa underkapitel.

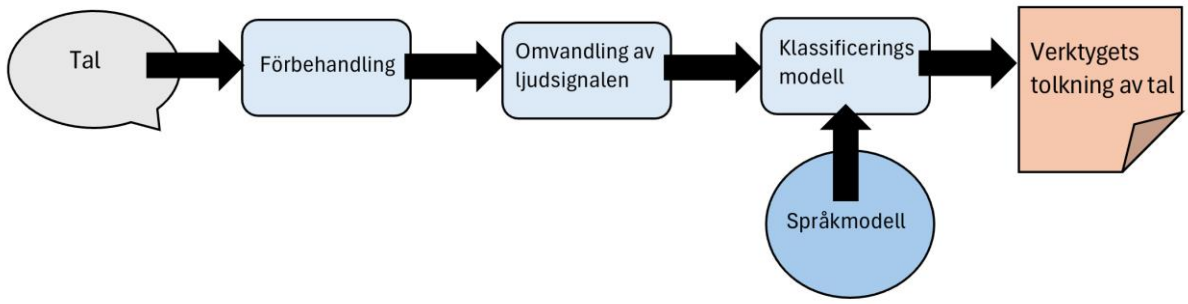
Även om vi har modeller som HMM är nyckeln till bättre taligenkänning att samla in mer data (dvs. tal som kan användas för att träna verktyg). Å andra sidan behöver man ha en tillräcklig bra algoritm att träna verktygen med för att vi ska kunna lösa alla problemen som variationen i talet orsakar. Som träningsdata för verktygen används det ofta uppläst tal, men O'Shaughnessy (2008:2966) påpekar att ASR framför allt bör kunna tolka spontant tal. Spontant tal har ännu mer variation i sig, eftersom talaren behöver tänka samtidigt som hen talar, och det således är aningen knepigare att tolka än uppläst tal. I nästa kapitel går jag närmare in på hur taligenkänning fungerar rent tekniskt.

4.1.1 Huvudprinciper i taligenkänning

Taligenkänningsverktyg omvandlar alltså tal till text, men vad är det som händer helt konkret? Det är det som vi ska ta itu i detta kapitel. Först presenteras de viktigaste komponenterna i taligenkänning och därefter tekniska krav och andra användbara komponenter. Slutligen säger jag några ord om taligenkänningsverktyget som används i denna avhandling.

Taligenkänning består av flera komponenter som tillsammans gör igenkänning av tal och dess bearbetning möjlig. Efter att vi har samlat och matat in tal eller ljud i ett verktyg ska ljudet först förbehandlas. Förbehandlingen görs främst för att förbättra ljudkvaliteten t.ex. genom att minska bakgrundsbuller. Häräst omvandlas den förbehandlade ljudsignalen till en mer informativ representation genom extraktion av den informationen som maskininlärningsmodeller behöver från ljudsignalen. Ljudsignalen omvandlas alltså till en form som är mer hanterbar för maskininlärningen (Gülbahar 2024) Den extraherade informationen går sedan genom en klassificeringsmodell. En klassificeringsmodell är en algoritm som används för att analysera och klassificera datan, dvs. den information som extraherades från ljudsignalen, på ett lämpligt sätt. HMM (*Hidden Markov Modell*), som

presenterades i föregående kapitel, kan användas som en klassificeringsmodell i traditionella taligenkänningsverktyg (Gülbahar 2024). En språkmodell är en viktig del här eftersom språkmodellen innehåller t.ex. de grammatiska reglerna och den semantiska informationen i ett språk. Således jobbar språkmodellen tillsammans med klassificeringsmodellen och gör korrigeringar enligt klassificeringsmodellens tolkning. (Papastratis 2021) Som ett resultat av dessa steg får vi en skriftlig tolkning av ljudet som vi utgick ifrån. Sedan kan man välja om man vill omarbeta eller granska det skriftliga resultatet på något sätt. I figur 1 illustreras dessa viktigaste komponenter i all enkelhet.



Figur 1. De viktigaste komponenterna i taligenkänning.

Det finns två viktiga tekniker som taligenkänningsverktyg fungerar med, den traditionella hybridmetoden och *end-to-end* -metoden. Den traditionella hybridmetoden kombinerar ett regelbaserat och ett statistiskt tillvägagångssätt. Med andra ord kombinerar metoden reglerna för varje språk med statistik över mönster som härrör från ett stort antal transkriberade ljudfiler. Detta är en effektiv metod, men beräkningsmässigt mycket krävande. Den andra metoden, den så kallade *end-to-end* -metoden, utnyttjar djupa neurala nätverk (DNN). Metoden utelämnar några tydliga mellansteg, som fonemigenkänning, vilket gör den till ett mer effektivt och till och med mer korrekt system. Denna metod är dock mer komplex samt kräver stora data- och beräkningsresurser för träning. (Sonix 2024)

För att kunna identifiera talare, olika accenter, dialekter och talmönster behöver taligenkänning använda sig av artificiell intelligens (AI) och maskininlärningsmodeller (*machine learning models*) (Cem 2024). Sonix (2024) konstaterar att det vid taligenkänning kan innebära t.ex. att AI känner igen den dialekt som talas och tar hänsyn till det vid bearbetning av talet. Maskininlärningsmodeller behövs däremot för att automatisera manuella och repetitiva maskininlärningsuppgifter. Enligt Cem (2023) kan automatiseringen påskynda processer, minska fel och ge mer exakta resultat. Tillsammans med AI och maskininlärning är

taligenkänning alltså utformad för att lära sig av varje interaktion och kontinuerligt förbättra prestandan och precisionen.

Andra komponenter som kan utnyttjas i taligenkänning är språkviktning, som betonar särskilda ord och fraser som ofta uttalas, t.ex. produktnamn. Detta gör det mer sannolikt att verktyget i fråga känner igen viktiga nyckelord. Sedan finns det en process för att filtrera och ta bort stötande eller olämpliga ord och fraser från ljuddata. Med akustiska modeller kan man fånga upp särskilda fonetiska enheter i ljudsignaler, alltså att avbilda ljudsignaler som morfem och fonem och således omvandla dem till digitala. Akustiska modeller tränas på stora datamängder som innehåller talprover från en mängd olika talare med olika accenter, talstilar och bakgrund. Vidare kan taligenkänning identifiera olika talare i ljudinspelningar genom att ge dem en unik identifikation. Detta underlättar talanalysen och identifiering av talare. (Gülbahar 2024)

Verktyget som jag använder för studien är utvecklat av ett finskt företag. Det kan känna igen tal antingen från ett diktat eller en fil. I den här studien består mitt material av ljudfiler och för varje ljudfil har jag valt att känna igen tal från en fil. Innan man matar in något i verktyget kan man välja vilken språkmodell man vill använda. Språkmodellen avgör vilket språk, vilken språkvariant eller till exempel vilket ämnesområde som stöds av taligenkänningsverktyget. Verktyget ger för sverigesvenskan två möjliga språkmodeller att välja mellan, en generell språkmodell och en språkmodell som är inriktad på politik. För finlandssvenskan anges dock endast en generell språkmodell. Eftersom jag ville att resultaten skulle vara så jämförbara som möjligt använde jag den generella språkmodellen för båda språkvarianterna. Taligenkänningsverktyget tillåter även att justering av andra inställningar, till exempel automatisk efterbehandling och korrigering av verktygets produktion. Jag har dock inte använt denna funktion, eftersom studien är inriktad på att undersöka de fel som produceras vid taligenkänning.

5 Taligenkänning som tillgänglighetsmedel

Utgångspunkten för utvecklingen av taligenkänningsverktyg var som framkommit att vi människor ville bli mer effektiva (Juang och Rabiner 2005:1–4) genom att utveckla verktyg som kan hjälpa oss bland annat att få ett jobb gjort snabbare (Jurafsky och James 2023, Forsberg 2003:2, Arora och Singh 2012:34). I kapitel 4.1 nämndes att kommunikation med röstkommandon är naturligt för människor (Engstrand 2012:20 f.). I det här kapitlet kommer jag att diskutera taligenkänning ur ett annat perspektiv. Taligenkänning är inte bara en kul grej som möjliggör att styra robotar med rösten, utan också ett nödvändigt hjälpmedel.

Samhällets digitalisering kommer att fortsätta och det är viktigt att säkerställa att alla medborgare kan använda digitala tjänster på lika villkor, trots att vi kan ha olika svårigheter som funktionsnedsättning eller svaga kunskaper i landets majoritetsspråk som finska i Finland. Det kan till exempel vara alltför svårt att förstora texten på skärmen eller webbsidor med komplicerad eller otydlig information kan göra det svårt att använda tjänsterna. Dessa problem är exempel på sådant som det nya tillgänglighetsdirektivet (EU) 2019/882) är avsett att förändra (Digin.nu).

Tillgänglighetsdirektivet träder i kraft år 2025, vilket betyder att det då kommer att ställas krav på hur digitala tjänster och produkter fungerar (Triggerfish.se). Det nya direktivet berör många aktörer, allt från elektroniska kommunikationstjänster och medietjänster till e-böcker, banktjänster och webbplatser och för alla finns det egna tillgänglighetskrav listade (Digin.nu). Det finns krav som gäller alla men även mer specifika krav för vissa branscher. Kraven ska uppfyllas den 28 juni 2025 och hinner man inte uppfylla dem kan det utgå sanktionsavgifter (ibid.). Alla detaljer finns beskrivna i *Europeiska unionens officiella tidning* på EU:s rättsdatabas (EUT L 151/70, 7.6.2019, s. 70). EU har dessutom ett intresse av att minska hinder mellan medlemsländer och med gemensam lagstiftning blir det också lättare att utveckla tillgängliga tjänster (Digin.nu).

Hjälpmedel är apparater eller program som underlättar livet för personer med funktionsnedsättning. En rullstol är en konkret exempel på ett hjälpmedel och i den digitala miljön kan hjälpmedel vara t.ex. datorprogram som läser texten på datorskärmen och förmedlar information till ljud. (Tillgänglighetskrav.fi) Man behöver inte ha en permanent sjukdom för att omfattas av tillgängligheten. Hjälpmedel och tillgängliga tjänster kan göra t.ex. e-handel lättare och snabbare för alla. Tillfälliga omständigheter, som graviditet eller en

bruten arm, kan också göra det svårt att använda en viss tjänst utan hjälpmedel (Gacek 2020). Tillgängliga tjänster bidrar till individens självständighet och ett mer inkluderande samhälle (Triggerfish.se).

Självbetjäningsterminaler, såsom kontakt- eller incheckningsautomater, är exempel på ställen för vilka det finns branschspecifika krav för tillgänglighet (Digin.nu). Självbetjäningsterminaler ska bland annat utnyttja text-till-tal-teknik. På detta sätt erbjuder taligenkänning ett sätt att förbättra samhällets tillgänglighet. Personer med motoriska begränsningar, till exempel ofrivilliga rörelser eller begränsade rörelsebanor, kan få särskild hjälp av att använda talstyrning i stället för ett tangentbord (Tillgänglighetskrav.fi). Vidare kan taligenkänningsverktyg användas för att texta ljud (Webbriktlinjer). Textning hjälper hörselskadade, men också språkinlärare och alla som just då inte vill eller kan titta på videor med ljud (Tillgängligashetskrav.fi). Ett mer tillgängligt samhälle ligger alltså verkligen i allas intresse.

6 Material och metod

I det här kapitel presenterar jag först det empiriska materialet och dess bearbetning närmare. Sedan i delkapitel 6.2 respektive 6.3 går jag genom såväl de kvantitativa som de kvalitativa metoderna som används i studien och motiverar varför båda metoderna behövs. Taligenkänningsverktyget som används i studien har presenterats kort tidigare i kapitel 4.1.1.

6.1 Material

Materialet i min avhandling är dels ljudfiler på två språkvarianter, finlandssvenska och sverigesvenska och dels taligenkänningsverktygets tolkning av dessa ljudfiler. Jag kommer nedan att gå igenom processen vid insamling av materialet, vad materialet består av och hur jag bearbetat materialet.

Trots att jag befinner mig i Finland visade det sig ganska fort att det är svårt att få tag på finlandssvenskt tal med någon sorts tillgänglig textning eller transkription. På sverigesvenska finns det mycket mer material att välja mellan, så jag fick helt enkelt att se vad jag hittar på finlandssvenska och anpassa studien efter det. Detta i sig visar maktförhållandet mellan finlandssvenska och sverigesvenska: på den ena varianten finns det hur mycket tillgängligt tal som helst och på den andra ganska lite. Dessutom ville jag gärna få ett lite mer begränsat ämnesområde som skulle vara gemensamt för båda språkvarianterna, vilket gjorde det ännu mer utmanande att hitta lämpligt material. Att samla tal själv är så klart alltid ett alternativ, men ett mycket tidskrävande sådant. Jag ville spara tid och därför ville jag få tag på tal som skulle kunna användas direkt, i stället för att hitta informanter och samla in tal från noll.

Politiska debatter finns tillgängliga på nätet och är till och med textade. Dessa diskussioner utgör mycket passande material, eftersom det förmodligen behandlas samma eller liknande frågor i finsk och svensk politik. Ett begränsat ämnesområde tillåter mig att se om taligenkänningsverktyget gör olika tolkning av precis samma ord och i fall talarens geografiska bakgrund har en påverkan. En annan orsak varför politiker kan tänkas vara en lämplig grupp av personer att testa taligenkänning med är att de ofta är väl förberedda talare. Reuter (2015b:57) beskriver vad som kännetecknar en bra offentlig talare: tydlig artikulation, att uttala ändelserna och slutkonsonanterna hörbart (som i *kastade*, *bordet* eller *talar*) samt att uttala korta ord som t.ex. *inte* och *skulle* som de skrivs. Politiker kan förväntas tillämpa dessa rekommendationer i sitt tal, alltså att tala skriftspråksnära. Vana talare som pratar mycket kan

också undvika suckar och stamning i sitt tal, det vill säga det kan bli mindre oklarheter för taligenkänningen att tolka. Av dessa anledningar valde jag riksdagsplenum från 2014 som finlandssvenskt material (Aalto-yliopisto, Aallon puheentunnistuskorpus eduskunnan istuntojen ruotsinkielisistä puheenvuoroista 2015-2020) och partiledardebatter från samma år som sverigesvenskt (Sveriges riksdag. Partiledardebatt 18 juni 2014.). Nedan presenterar jag smakprov ur materialet för att ge en idé av vilken typ av tal ingår i materialet.

(1) Ur finlandssvenskt material

> Talman! Det finns ingen orsak att förlänga den här debatten längre än nödvändigt. Det mesta har redan sagts. Ändå finns det skäl att konstatera att Finlands riksdag på fredagen kan fatta ett historiskt beslut. Efter att den föreslagna ändringen i äktenskapslagen samt de andra lagändringar som regeringen med samma riksdagsbeslut uppmanas bereda har kommit i kraft 2017 intar Finland sin plats i den nordiska värdegemenskap där vi till den här delen hittills har gömt oss i skåpet. Ett viktigt signalement för rättsstaten är allas likhet inför lagen samt samhällets starka signal om att alla individer är lika värdefulla och accepterade. Så har det inte varit hittills, eftersom samhället i praktiken har graderat kärleksrelationer mellan två vuxna personer såsom juridiskt starkare och svagare enbart beroende på parens kön. Finland är ett bra land. Efter omröstningen på fredagen är Finland en lite bättre plats att leva i.

I exempel 1 talar den finska riksdagsledamoten för den nya äktenskapslagen som trädde i kraft i Finland den 1 mars 2017.

(2) Ur sverigesvenskt material

> Herr talman! Det problematiska med Gustav Fridolin – det gäller inte bara detta ämne utan alla ämnen – är att han alltid ska göra allt. Han ska alltid göra både och. Det saknas aldrig resurser. Det är bara att gratulera Miljöpartiet till att ha hittat ett språkrör som har så mycket pengar i sina fickor att han kan spendera precis allting på precis allting. Men så fungerar inte verkligheten. I verkligheten har man en begränsad mängd pengar. Vi har politiker som väljs av folket och som bestämmer hur vi ska prioritera och fördela pengarna. Det är ganska grundläggande. Ofta vill man göra mer än en sak, och då har vi olika verksamheter som vi fördelar pengarna mellan. Man kan inte ge så mycket pengar som behövs till alla verksamheter, utan man måste prioritera. Då menar jag att det

> år mer effektivt att hjälpa de människor som är flyktingar där de allra flesta finns och där kostnaderna för hjälpen är väsentligt lägre än i till exempel Sverige. Min politik är mycket mer effektiv än din, Gustav Fridolin. Så enkelt är det.

I det sverigesvenska materialet är fokus lite mer personligt. Det tas upp namn på sådana som har lovat göra ett och annat och diskuteras sedan hur det egentligen har gått för dem med att hålla sin löften. Kritik tycks alltså ges åt personer snarare än åt åtgärder och saker.

Riksdagsplenum från Finland innehåller förstås mycket tal på finska. Ibland vill finskspråkiga ledamöter delta i någon diskussion på svenska eller så kan det vara med svenskspråkiga som är uppvuxna i Sverige. För att det finlandssvenska materialet ska vara autentiskt kunde jag inte ta med alla som pratar någon sorts svenska. Det är viktigt att materialet består av modersmålstalare så att språkkunskapen är på en sådan nivå att Reuters översikt över uttalsdrag kan tillämpas (se kapitel 2.1.1). Jag kunde inte riskera att taligenkänningen skulle göra felaktiga tolkningar på grundval av att talaren uttalar felaktigt. Det här betydde att jag behövde klippa och klistra ljudfiler från riksdagsplenumet en hel del för att endast få med finlandssvenska eller tvåspråkiga talare i materialet från Finland. Vad gäller partiledardebatter från Sverige var det lite mer okomplicerat eftersom jag kunde utgå från att alla är svenskspråkiga modersmålstalare. Vid bearbetningen av materialet använde jag datorprogrammet *Audacity*.

Det tog alltså mycket längre att tid bearbeta det finlandssvenska materialet. Jag gick genom alla riksdagsplenum under en sexmånadersperiod och klippte systematiskt ut alla talturer på finlandssvenska. Jag klistrade sedan ljudklippen ihop så att varje ljudfil representerar en dag och innehåller alla talturer från den dagen. En ljudfil kan således bestå av en eller flera taltur respektive talare. För det finlandssvenska materialet blev det 30 ljudfiler vilket betyder ungefär 88 minuter tal och totalt 10 932 ord.

Det finlandssvenska materialet består alltså av talturer i plenum under en sexmånadersperiod, medan det sverigesvenska materialet består av samtal under en och samma dag. Det sverigesvenska materialet var egentligen färdigt att mata in i taligenkänningsverktyget som sådant, men jag klippte till det till kortare repliker för att det skulle vara snabbare för verktyget att hantera. I det sverigesvenska materialet fortsätter talare ofta med sina repliker direkt efter varandra, vilket sällan händer i det finlandssvenska materialet där det kan vara att det just den dagen inte finns någon som kan reagera på en taltur på finlandssvenska. För det

sverigesvenska materialet blev det också 30 ljudfiler men bara ungefär 68 minuter tal och totalt 11 046 ord.

6.2 WER (*word error rate*)

WER (*word error rate*) är en standardiserad metod för att mäta och betygsätta taligenkänningsverktyg kvantitativt. Då man går genom det som taligenkänning har tolkat och jämför det med en ursprunglig transkription kan man se alla gånger där taligenkänningen har gjort en fel tolkning av talet. Alla felen räknas ihop och antalet fel delas med det totala antalet ord i transkriptionen så att vi får en WER-procentandel. Det finns tre typer av fel, strykning, tillägg och ersättning, vilka alla kan analyseras närmare. Nedan ges ett kort exempel på varje felkategori:

(3) […] vill köpa inhemska *och* närproducerade livsmedel.
[…] vill köpa inhemska närproducerade livsmedel.

(4) Vi vet att konsumenterna […]
Vi *har* vet att konsumenterna […]

(5) *Värderade* talman! Enligt statistikuppgifter har systemet […]
Meddelade talman! Enligt statistikuppgifter har systemet […]

I exempel 3 har det skett en strykning av ordet *och*, i exempel 4 har ordet *har* lagts till och i exempel 5 finns det en ersättning av ordet *värderade*. WER anger alltså procentandelen fel, t.ex. i exempel 4 är WER-procenten 0,25 % eftersom det finns ett fel och fyra ord totalt (ett delad med fyra). Således är andelen av korrekt tolkning 75 % i exempel 4.

Även om WER används i stor utsträckning finns det vissa svagheter. Alla fel i en och samma felkategori är inte nödvändigtvis lika allvarliga eller sinsemellan likadana. Felkategori ersättning innehåller till exempel både fel där ett ord har fått en felaktig böjningsform (*huset* har tolkats som *husen*) och fel där ett ord har blivit något helt annat, som i exempel 5 ovan, vilket kan förändra budskapet totalt. Därför är det viktigt att gå genom och analysera felkategorierna vidare, samt göra en mer djupare, kvalitativ analys om felen.

Eftersom identifiering av felen och beräkning av WER-procenten är långtifrån en entydig metod ville jag identifiera och räkna felen manuellt, dvs. själv gå igenom ljudfiler, transkriptioner och verktygets tolkningar, hellre än programmera det. Detta skulle ändå krävts för att göra kvalitativ analys, men att jobba på det sättet säkerställer också att jag är har

kontrollen över hur och vilka fel som identifieras. En manuell analys av WER tillåter mig att titta på transkriptionen och lyssna på ljudfilerna samtidigt då jag gör analysen, i fall det skulle vara fall där felet är egentligen hos transkriptionen snarare än i taligenkänningsverktygets tolkning. Här måste man alltså ta hänsyn till att det finns risk för mänskliga fel, ett fel kan gå obemärkt förbi eller hamna i fel kategori. Det finns till och med en del oklara fall där felen skulle kunna placeras i flera kategorier eller räknas på flera olika sätt. Ett exempel på detta illustreras nedan.

(6) […] den dömda kan placeras direkt i ett fängelse med en lämplig sysselsättningsverksamhet och säkerhetsnivå.
[…] den *dömde kontrolleras* direkt i ett fängelse med en lämplig sysselsättningsverksamhet och säkerhetsnivå.

Det första felet i exempel 6 är enkelt att identifiera som ersättning (*dömda* har blivit *dömde*), men övriga felen är svårare att identifiera. De kan antigen identifieras som två strykningar (*kan* och *placeras* har strukits) och ett tillägg (*kontrolleras* har lagts till). En annan möjlig tolkning är att identifiera felen som en strykning (ordet *kan*) och en ersättning (ordet *placeras* ersättas med *kontrolleras*). Jag har dock utgått från transkriptionen och identifierat detta därför som det sistnämnda.

Vi vet mycket om vad som orsakar problem vid taligenkänning (se kapitel 4.1), men trots detta anger inte WER-procenten orsaken varför ett visst fel uppstår. För att kunna försöka förstå orsakerna bakom felen behöver vi en djupare analys som också kan leda oss till förbättringsförslag för verktygen. Verklig förståelse av talat språk kräver mer än hög noggrannhet i att känna igen enskilda ord, både för människor och för taligenkänning (Wang, Acero och Chelba 2003:583). Enligt Kakkar (2023) är 0,15 % eller lägre redan en mycket bra WER-procent.

6.3 Komparativ metod

Som det konstaterades i föregående kapitel är WER-procenten inte ett helt entydigt sätt att betygsätta taligenkänning och därför vill jag även analysera materialet kvalitativt med en komparativ metod. En mer kvalitativ analys kompletterar den kvantitativa delen och ger en möjlighet att utreda varför de kategoriserade felen uppstår.

Jag gör en komparativ jämförelse i två delar. Först jämför jag taligenkänningsverktygets tolkning med de ursprungliga transkriptioner, detta görs i samband med att identifiera verktygets feltolkningar för WER-procenten. Att djupare analysera hurdana fel det finns i de olika kategorierna och hur stora felkategorierna är är till hjälp i nästa steg där jag riktar uppmärksamheten mot de två språkvarianterna.

Efter att jag har samlat och kategoriserat felen för var sin språkvariant tittar jag närmare på i felkategorierna och jämför språkvarianterna med varandra. Då är det fokuset på om verktyget har gjort liknande tolkningsfel för båda språkvarianterna eller inte. Här utgår jag framför allt från Reuters lista över de avvikande uttalsdrag som jag presenterade i kapitel 2.1.1.

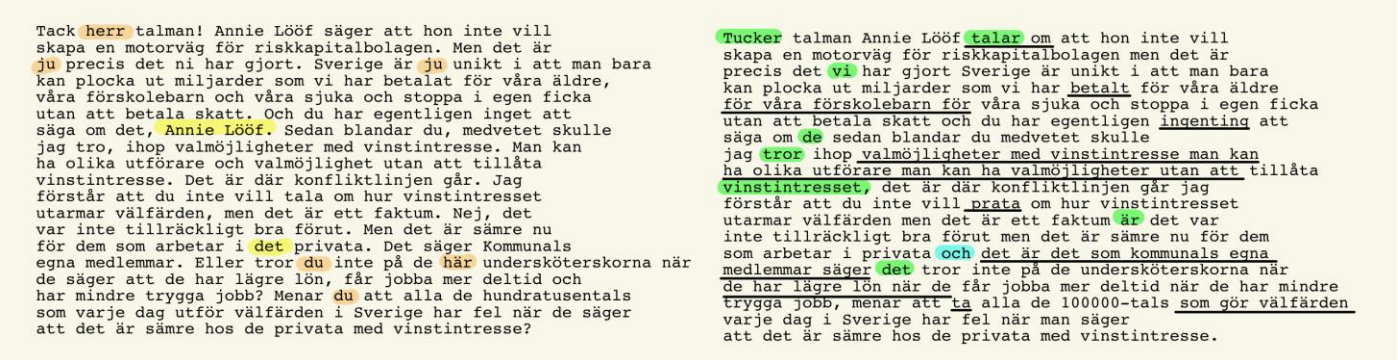
7 Finlandssvenska och sverigesvenska ur taligenkänningsverktygets feltolkningar

I detta kapitel går jag igenom resultaten ur kvantitativt synvinkel. Jag inleder kapitlet med att presentera de utmaningar som jag stöttade på vid analysen. Därefter ger jag en översikt över de tre felkategorierna som WER bygger på och går sedan igenom kategorierna mer detaljerad. Kapitel 8 utgör analysen ur kvalitativt synvinkel och behandlar orsak som eventuellt har lett till dessa feltolkningar.

Den största utmaningen var att transkriptioner som kom med tal visade sig vara ganska grova. Fall där talaren korrigerar sig själv, använder lite av en ledigare register, preciserar uttalet eller talar i en ogrammatisk ordning har ofta förenklats och korrigerats i transkriptionen. Korrigeringar gör att man över huvud taget kan följa med talet. Som Engstrand (2012:9) skriver är tal ofta ogrammatiskt och fullt med avhuggna meningar. Detta leder till avvikelser mellan transkriptioner och det som talaren egentligen säger i form av ordföljd och några enstaka ord, vilket betyder att jag inte kunde blind lita på transkriptioner. Jag kunde alltså inte räkna alla ställen där taligenkänningsverktygets tolkning avvek från transkriptionen som fel. I detta fall är vissa omskrivningar dock helt motiverade. Vi bör komma ihåg vilken funktion en transkription fyller, meningen är att den fungerar som hjälpmedel för personer som inte kan lyssna på debatten. Att följa tal i en skriven form exakt så som det sägs är inte särskilt lätt och därför har transkriptionen omarbetats en aning.

Av denna orsak var jag många gånger tvungen att lita på mina öron, snarare än transkriptionen. Det var också en orsak varför räkningen av WER-procenten kunde inte automatiseras. Fall där avvikelser mellan taligenkänningsverktygets tolkning och den ursprungliga transkriptionen orsakades egentligen av omskrivningar i transkriptionen kan inte räknas vara fel hos taligenkänningen. Ett exempel på hur komplext det kunde se ut illustreras i exempel 7 nedan. Till vänster finns den ursprungliga transkriptionen och till höger taligenkänningsverktygets tolkning.

(7) Exempel på fall där avvikelser i transkriptionen och taligenkänningsverktygets tolkning orsakas av omskrivningar i transkriptionen



Tack herr talman! Annie Lööf säger att hon inte vill skapa en motorväg för riskkapitalbolagen. Men det är ju precis det ni har gjort. Sverige är ju unikt i att man bara kan plocka ut miljarder som vi har betalat för våra äldre, våra förskolebarn och våra sjuka och stoppa i egen ficka utan att betala skatt. Och du har egentligen inget att säga om det, Annie Lööf. Sedan blandar du, medvetet skulle jag tro, ihop valmöjligheter med vinstintresse. Man kan ha olika utförare och valmöjlighet utan att tillåta vinstintresse. Det är där konfliktlinjen går. Jag förstår att du inte vill tala om hur vinstintresset utarmar välfärden, men det är ett faktum. Nej, det var inte tillräckligt bra förut. Men det är sämre nu för dem som arbetar i det privata. Det säger Kommunals egna medlemmar. Eller tror du inte på de här undersköterskorna när de säger att de har lägre lön, får jobba mer deltid och har mindre trygga jobb? Menar du att alla de hundratusentals som varje dag utför välfärden i Sverige har fel när de säger att det är sämre hos de privata med vinstintresse?

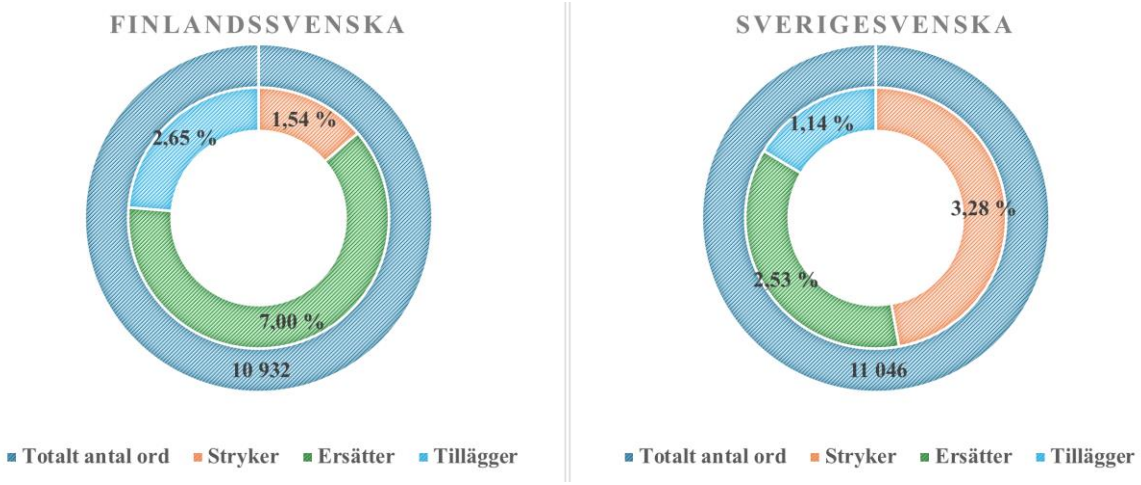
Tucker talman Annie Lööf talar om att hon inte vill skapa en motorväg för riskkapitalbolagen men det är precis det vi har gjort Sverige är unikt i att man bara kan plocka ut miljarder som vi har betalt för våra äldre för våra förskolebarn för våra sjuka och stoppa i egen ficka utan att betala skatt och du har egentligen ingenting att säga om de sedan blandar du medvetet skulle jag tror ihop valmöjligheter med vinstintresse man kan ha olika utförare man kan ha valmöjligheter utan att tillåta vinstintresset, det är där konfliktlinjen går jag förstår att du inte vill prata om hur vinstintresset utarmar välfärden men det är ett faktum är det var inte tillräckligt bra förut men det är sämre nu för dem som arbetar i privata och det är det som kommunals egna medlemmar säger det tror inte på de undersköterskorna när de har lägre lön när de får jobba mer deltid när de har mindre trygga jobb, menar att ta alla de 100000-tals som gör välfärden varje dag i Sverige har fel när man säger att det är sämre hos de privata med vinstintresse.

I exempel 7 har jag markerat feltolkningarna strykning, ersättning och tillägg med respektive färger orange, grön och blå. Med gul färg har jag i transkriptionen markerat ord som talaren faktiskt inte säger (*Annie*, *Lööf* och *det*). Dessa har antagligen tillagts i transkriptionen för att hjälpa läsaren att hänga med. Dessa ord har jag inte räknat med i totalt antal ord. Ytterligare har jag i verktygets tolkning understrukit ställen som taligenkänning har tolkat helt korrekt men som avviker från transkriptionen på grund av gjorda omskrivningar. På första raden ser vi till exempel att talaren säger ”Annie Lööf *talar om* att hon inte vill skapa en motorväg” vilket har i transkriptionen skrivits om till ”Annie Lööf *säger* att hon inte vill skapa motorväg”. På vissa ställen har det i transkriptionen använts precis samma ord som talaren använder men i en annan ordning.

Detta exempel (7) kan vara lite utmanande att tolka för läsaren, men det som jag vill lyfta fram är mängden understrykningar, dvs. mängden omskrivningar i transkriptionen. Dessa omskrivningar är naturligtvis inte fel hos taligenkänning och påverkar inte heller verktygets tolkning av talet, således är de inte heller färgkodade med någon felkategori. Detta förekommer dock i så stor utsträckning att jag måste aktivt ta det i hänsyn i analysen för att minimera risken för felens feltolkningar. Därför var det nödvändigt att lyssna på ljudfilerna samtidigt då jag jämförde taligenkänningsverktygets tolkning med transkriptioner. Det är också värt att notera att taligenkänningsverktyget i fråga inte identifierar skiljetecken om man inte uttalar dem i sitt tal, därför har t.ex. frågetecknet i slutet av exempel 7 inte kommit med i verktygets tolkning. Detta är en funktion, inte en bugg.

Näst följer en kvantitativ genomgång av materialet utifrån de tre felkategorierna ersättning, strykning och tillägg, som presenterades närmare i kapitel 6.2. En helhetsbild över

felkategorierna illustreras i figur 2. Jag visar sedan exempel med lägsta och högsta felprocenten i båda språkvarianterna. Efter det går jag mer detaljerad genom var sin felkategori och huvudsakligen om taligenkänning har gjort liknande feltolkningar i språkvarianterna.



Figur 2. Felkategorierna som en procentandel av det totala antalet ord i var sin språkvariant.

De yttersta ringarna i figur 2 visar det totala antalet ord i materialet och de inre ringarna i representerar de tre felkategorierna och visar andelen fel i materialet som helhet. Det kan läsas i figuren att i det finlandssvenska material den överlägset största felkategorin är ersättningar med exakt sju procent medan de två andra felkategorierna, strykningar och tillägg, är sammanlagt under fem procent. Däremot i det sverigesvenska material finns det två större felkategorier, nämligen strykningar med lite över tre procent och ersättningar med ungefär två och en halv procent, och tillägg är den överläget minsta kategorin med lite över en procent.

Det finns tre ljudfiler i det finlandssvenska materialet som delade den lägsta WER-procenten, vilket var 0,07 %, dessa exempel innehöll alltså minst fel. Nedan presenteras ett av dessa i sin helhet, först den ursprungliga transkriptionen där jag har markerat strukna ord och sedan taligenkänningsverktygets tolkning där jag har markerat ersätta och tillagda ord. Strukna ord markeras med orange, ersättningar med grön och tilläggningar med blå, färgerna är således desamma som i exempel 7 och figur 2 ovan. Jag använder även symbol * för att markera stället där transkriptionen byter till taligenkänningsverktygets tolkning, detta hjälper framför allt att följa med längre exempel framöver. Dessa färgkoder och layout tillämpar jag i alla framtida exempel.

(8) Lägsta WER-procent i finlandssvenskt material

Värderade talman! Det är viktigt att Finland håller fast vid löftet som vi har gett om att reservera 0,7 % av vår BNP i u-landssamarbete. Behovet av hjälp ute i världen minskar inte, det ökar. De facto har läget inte blivit bättre i de allra fattigaste länderna i världen och vi vet att det råder humanitära kriser på många, många håll. I slutet av nästa år ordnas den stora klimatkonferensen i Paris och i samband med det försöker man också få ihop pengar för en grön klimatfond. Kommer Finland att kunna leva upp till sitt löfte om att delta aktivt i den här fonden, och på det viset bidra till att hjälpa de fattigaste länderna i klimatproblematiken?

* Vardera tar man! Det är viktigt att Finland hålla fast vid löftet som vi har gett om att reservera 0,7 % av vår BNP i Ålands samarbete behovet ute i världen minskar inte av hjälp utan det ökar de facto har läget inte blivit bättre i det mest i de allra fattigaste länderna i världen och vi vet att det råder humanitära kriser på många många håll i slutet av nästa år ordnas den stora klimatkonferensen i Paris och i samband med det försöker man också få ihop pengar för en grön klimatfond kommer Finland att kunna leva upp till sitt löfte om att delta aktivt i den här fonden och både vice bidra till att hjälpa de fattigaste länderna i klimatproblematiken.

I exempel 8 ovan ser vi att taligenkänningsverktygets tolkning avviker från transkriptionen på rad två och tre. Detta är ytterligare ett exempel på det som jag tidigare illustrerade med exempel 7, nämligen att avvikelserna orsakas av omskrivningar i den ursprungliga transkriptionen. På rad två och tre uttrycker talaren sig på ett väldigt ledigt sätt, *behovet ute i världen minskar inte av hjälp utan det ökar*, och lite senare korrigerar hen sig själv, *har läget inte blivit bättre i det mest [kort paus] i de allra fattigaste länderna*. Dessa ställen har redigerats i transkriptionen. För tydlighetens skull kommer jag dock inte längre markera sådana avvikelser i exempel, utan om de inte har färgkodats som feltolkningar är det korrekt i taligenkänningens tolkning.

Det kan snabbt läsas från exempel 8 att det finns väldigt lite färgmarkeringar och således feltolkningar. Det finns inga tillagda ord och endast ett ord har blivit struken av taligenkänningsverktyget (*det*). Det kan diskuteras vilka ord egentligen har blivit strukna och vilka ersatta (där *på det* har tolkats som *både*), men en annorlunda tolkning i detta fall skulle inte påverka antalet fel. Som figur 2 tyder är ersättningar det vanligaste felet som

förekommer. För jämförelse presenteras exemplet som fick den högsta WER-procenten 0,42 %, dvs. innehöll mest fel i det finlandssvenska materialet.

(9) Högsta WER-procent i finlandssvenskt material

Värderade talman! I slutet av varje valperiod hör det till att ministerierna skriver framtidsöversikter som sedan utgör en del av bakgrundsmaterialet på basis av vilket följande regeringsprogram skrivs.

* All man i stället var det var perioden hörde till att ministeriernas skrivare framtidsöversikter som sedan utgör en del av bakgrundsmaterialet på basis av vilka följande regeringsprogram skrivs.

Exempel 9 representerar en mycket kort ljudfil men det ingår också mycket mera fel i den. Även om *talman* är en sammansättning har jag identifierat felet så att endast ordet *tal* har ersätts med ordet *all*. Den här typen av fel inträffar mycket ofta och även i föregående exempel 8. Vidare har en annan sammansättning, ordet *valperiod,* tolkats som två fel, *var* och *perioden*, eftersom det sista ordet har fått en fel ändelse. Det som framgår i figur 2 stämmer i exempel 9 också, det är betydligt klart att ersättningar utgör majoriteten av fel, sedan finns det tre strykningar och ett tillägg (*det* med blå).

Vi tar motsvarande exempel också ur det sverigesvenska materialet. Det allra lägsta WER-procent bland det sverigesvenska materialet var 0,03 % och det var så många som fem ljudfiler som delade denna procent. Ett av dessa illustreras nedan i sin helhet.

(10) Lägsta WER-procent i sverigesvenskt material

Tack herr talman! Kriget i Syrien går in på sitt tredje år. Kemiska vapen har använts, och 10 000 barn beräknas vara döda. Vittnen har rapporterat att barn, spädbarn och gravida utsätts för krypskyttar, summariska avrättningar och tortyr. 1,2 miljoner barn är på flykt. Ungefär hälften av dem är under fem år, och många är skilda från sina föräldrar. Ytterligare en miljon barn som är kvar i Syrien lever i städer som närmast är under belägring. De är utan mat och vatten och har inte medicin ens mot de enklaste sjukdomar. Och nu sprids kriget i Irak. Sverigedemokraternas vidlyftiga löften i debatten baseras på pengar som ni, Jimmie Åkesson, låtsas kunna dra in genom att säga nej till nio av tio som kommer till Sverige och söker en fristad. Hittills har ni inte svarat på vilka ni ska skicka tillbaka till krigets helvete, men nu säger Sverigedemokraterna att vi inte

ska ge uppehållstillstånd till flyktingar från Syrien. Asylsystemet finns inte av en tillfällighet. Efter andra världskrigets fasor och första världskrigets grusade förhoppningar när det gäller att kunna stoppa framtida krig utfäster vi i alla fall detta löfte till varandra: Om någon av oss i mänsklighetens familj utsätts för krigets helvete igen då ska vi kunna lita till rätten att fly och få en fristad i ett annat land. I FN:s allmänna deklaration om de mänskliga rättigheterna uttrycks detta så här: ”Envar har rätt att i annat land söka och åtnjuta fristad från förföljelse.” Låt mig fråga Jimmie Åkesson detta: Om Sverigedemokraterna inte anser att det krig där barn nu dödas och lemlästas i Syrien är av sådant slag som avsägs när deklarationen om de mänskliga rättigheterna undertecknas, vad krävs då för att en flykting ska betraktas som en flykting med mänskliga rättigheter också av Sverigedemokraterna?

* Tacka talman och kriget i Syrien går in på sitt tredje år kemiska vapen används 10 000 barn beräknas stöda vittnen har rapporterat att barn och spädbarn gravida utsätts för krypskyttar summariska avrättningar och tortyr i 1,2 miljoner barn är på flykt ungefär hälften av dem under fem år. Och många enskilda från sina föräldrar ytterligare en miljon barn som är kvar i Syrien lever i städer som närmast är under belägring utan mat, vatten eller ens medicin mot de enklaste sjukdomar och nu sprids kriget i Irak Sverigedemokraternas vidlyftiga löften i den här debatten baseras på pengar ni låtsas kunna dra in genom att säga till nio av 10 av dem som kommer till Sverige och söker en fristad nej hitintills har ni inte svarat på vilka det ni ska skicka tillbaka till krigets helvete men nu säger Sverigedemokraterna. Att de flyktingar från Syrien ska vi inte ge uppehållstillstånd och asylsystemet finns inte av en tillfällighet efter andra världskrigets fasor och de grusade förhoppningarna från första världskriget om att inte kunna stoppa framtida krig utfäster vi i vart fall detta löfte till varandra om någon av oss i mänsklighetens familj igen utsätts för krigets helvete ska vi kunna lita till rätten att fly och få en fristad i ett annat land i FN:s allmänna deklaration om de mänskliga rättigheterna uttrycks detta alla har rätt att i annat land söka och åtnjuta fristad från förföljelse, låt mig fråga Jimmie Åkesson om Sverigedemokraterna inte anser att det är krig där nu barn dödas och lemlästas i Syrien är av sådant slag som avsägs när deklarationen om mänskliga rättigheterna underteckna vad krävs då för att en flykting ska betraktas som en flykting med mänskliga rättigheter också av Sverigedemokraterna.

De sverigesvenska exemplen är lite längre som vi ser här i exempel 10. På liknande sett som med exempel 8 ser man fort att här förekommer det väldigt få feltolkningar, vilket kan förväntas då felprocenten är så förvånansvärt låg. Figur 2 visar att den största felkategorin i det sverigesvenska materialet är strykningar och näst största ersättningar. Just i detta exempel är det dock tvärtom även om det handlar om små skillnader, fyra ersättningar och tre strykningar. Tilläggningar är den minsta kategorin i det sverigesvenska materialet, här i exempel 10 med två förekomster.

Exempel 10 ovan är också ett bra bevis på det som konstaterades i början av kapitlet, att det är inte så enkelt att läsa talat språk och försöka hänga med på alla ställen i verktygets tolkning. Talaren säger till exempel *ni låtsas kunna dra in genom att säga till nio av 10 av dem som kommer till Sverige och söker en fristad nej*. Detta har i transkriptionen förenklats med att flytta ordet *nej* tidigare, sådana här omskrivningar är alltså nästan nödvändiga för att transkriptionen skulle vara till hjälp för till exempel hörselskadade personer. I exempel 10 har jag ytterligare gulmarkerat ord som har lagts till i transkriptionen, alltså ord som talaren har inte sagt men som fungerar som stödmedel för läsaren. Här vill jag lyfta fram att sådana här omskrivningar förekommer särskilt mycket i det sverigesvenska materialet, men jag kommer inte att markera dessa ord i fortsatta exempel. Jag påminner att de gulmarkerade orden har uteslutits från det totala antalet ord och påverkar således inte WER-procenten.

Utelämnade skiljetecken kan också göra det svårare att läsa exempel men jag har inte viljat redigera taligenkänningsverktygets tolkning på det sättet. Å andra sidan har taligenkänningen ibland tolkat att det t.ex. finns en punkt där det egentligen inte gör det eftersom början och slut på meningar signaleras på ett annat sätt i tal än i text. Näst presenteras exemplet med den högsta WER-procenten 0,2 % ur det sverigesvenska materialet, alltså med mest feltolkningar.

(11) Högsta WER-procent i sverigesvenskt material

Herr talman! Jonas Sjöstedt sade att Socialdemokraterna har ett val om de rödgröna får majoritet, nämligen att regera åt antingen höger eller vänster. Jag antar att ”höger” betyder jag eller Annie Lööf – eller Göran Hägglund kanske vad vet jag? Kanske skulle Fredrik Reinfeldt ingå. Göran Hägglund såg lite frestad ut en stund när han hörde att Sjöstedt skulle vara med, men sedan ångrade han sig. Sanningen är att jag kan lugna Jonas Sjöstedt. Du får ha Socialdemokraterna för dig själv. Du behöver inte vara orolig; ingen av oss kommer att sätta sig i en regering med de rödgröna om det skulle bli ett sådant flertal. Jag kommer att göra

allt jag kan för att det ska undvikas, men om svenska folket röstar så då får Vänsterpartiet vara med och styra. Jag lovar. Ni får vara med. Ett besked har vi ju fått, inte bara här i dag utan framför allt de senaste dagarna, och det är att Vänsterpartiet tycker att man inte ska ha läxor i skolan. Jag håller alltså med; ni ger en del besked som är viktiga. Det är ingen liten fråga, utan det handlar om hur skolan ska arbeta och om våra undervisningsresultat i Sverige. Det handlar om att kunskapsresultaten har sjunkit under 20 års tid i svensk skola. Jag tror att en av de åtgärder som måste vidtas för att lyfta resultaten det är att elever behöver anstränga sig mer. Det är inte den enda åtgärden – vi behöver en bättre lärarutbildning, mer resurser och allt möjligt – men elever måste anstränga sig mer, och det måste vi våga säga. Jag fattar ju att skolungdomar som intervjuas säger: Läxfritt? Jippie, då kan jag göra annat! Vi vuxna har dock ett lite större ansvar, och vi måste våga säga detta. Då säger Vänsterpartiets ledande skolpolitiker – jag hade en debatt häromdagen – att det inte är läxor ni är emot. Läxorna ska dock få göras i skolan, så det är hemläxor vi är emot säger man då. Det är alltså något försök till nyspråk – jag trodde liksom att läxa var hemläxa. Annars kallar vi det lektioner. Men liksom det, det är, men ja berätta gärna varför ni vill stoppa hemläxor i den svenska skolan, Jonas Sjöstedt.

* Herr talman är Jonas Sjöstedt sade att socialdemokraterna har ett val om de rödgröna få majoritet eller flertal att antingen göra åt höger eller vänster antar att det här med högre betydde att det var jag eller Annie Lööf / Göran Hägglund kanske varit jag eller Fredrik Reinfeldt skulle ingå Göran Hägglund såg lite frestad ut en stund när han hörde att Sjöstedt skulle vara med ändå sedan ångra sig är sanningen är att jag kan göra det kan lugna Jonas Sjöstedt du får ha socialdemokraterna för dig själv det du behövde var orolig, det är ingen av oss kommer att sätta oss i någon regering med de rödgröna om det skulle bli ett sådant flertalet om svenska folket röstar så jag kommer att göra allt vi kan för att det ska undvikas med de svenska folket röstar så får vänsterpartiet är ett besked vi fått här i dag men framför allt de senaste dagarna och vänsterpartiet tycker att man inte ska ha läxor i skolan så jag håller med en egen del besked som är viktiga är ingen liten fråga det handlar om hur skolan ska arbetas det handla om våra undervisningsresultat i Sverige handlar om att kunskapsövertagarna sjunkit under 20 års tid i svensk skola jag tror att en av de åtgärder måste vidtas för att lyfta resultat är att elever behöver anstränga sig mer, jag tror det är inte den enda

> åtgärden vi behöver bättre lärarutbildning mer resurser allt möjligt men också elever måste anstränga sig och det måste vi våga säga att jag fattar när man ser skolungdomar intervjuar sig om läxfritt gips kan jag göra annat men vi som vuxna har ett lite större ansvar, vi måste våga säga och då säger vänsterpartiet Årets ledande skolpolitiker hade debatt häromdagen hade inte växer vi mot Leksand ska däremot får göras i skolan är det hemläxor vid motsäger man försökte nyspråk att växa vår hemläxa annars kallade för lektioner men liksom det d men jag berätta gärna Jonas Sjöstedt var varför vill ni stoppa hemläxor i den svenska skolan.

Det första ordet *är* som har lagts till är ett intressant fall. Där säger talaren alltså egentligen ingenting, därför har jag felmarkerat ordet som ett tillägg. Talaren tänker tydligt på vad hen ska säga och tvekar. Det finns alltså en paus i talet, men den fylls inte av tystnad eller inandning, utan av ett slags eftertryckligt uttal som liknar r-uttal, därför har verktyget tolkat att ordet sannolikt är *är*. Den mänskliga hjärnan skulle lätt ha kunnat göra en liknande feltolkning. Men eftersom uttalet där snarare är en tankepaus eller ett sätt att strukturera talet och inte ett ord som skulle vara nödvändigt för lyssnaren att registrera, måste det räknas som en feltolkning. Ett annat intressant fall i verktygets tolkning är ett snedstreck i *jag eller Annie Lööf / Göran Hägglund*. Det är möjligt att ange kommandon till exempel för olika förkortningar i taligenkänningsverktyget, men här kan jag inte vara säker på vad exakt verktyget har tolkat som ett snedstreck. Därför har jag i detta fall identifierat ordet *eller* som struken. Detta kan möjligen läsas som ett falskt fel, men med mängden fel i exemplet påverkar inte ett enda fel nämnvärt på WER-procenten.

Om vi gör en liten sammanfattning av dessa fyra exempel ser vi att de två exempel med lägsta WER-procent, dvs. exempel 8 och 10, ligger under tröskeln på 0,15 % som enligt Kakkar (2023) är gränsen för ett bra resultat. Rent numeriskt sett räknas dessa två exempel att vara lättbegripliga. Det finns dock en allvarligare fel som ändrar budskapet i båda exemplen. I det finlandssvenska exemplet har *u-landssamarbete* blivit *Ålands samarbete*. I det sverigesvenska exemplet står det att *10 000 barn beräknas stöda* i stället för *döda*.

Till skillnad från de bäst presterande exemplen, som tydligt passerade tröskelvärdet, går exempel 9 och 11 tydligt över gränsen. Eftersom exempel 9 har den högsta WER-procenten ur finlandssvenskt material och felprocenten är till och med högre än i det sverigesvenska exemplet 11, betyder det att exempel 9 innehåller sammanlagt mest fel ur hela materialet. Fördelningen av felkategorier i exempel 9 följer samma mönster som visas i figur 2 och

exempel 8. Taligenkänningens feltolkningar är koncentrerade till början av exemplet, vilket gör tolkningen vagt och otydligt. Det kommer t.ex. inte fram vad hen egentligen talar om då *varje valperiod* har tolkats som *var det var perioden*. Trots att det finlandssvenska exemplet 9 har högre felprocent än det sverigesvenska exemplet 11 blir det sverigesvenska kanske även flummigare. Fördelning av felkategorier i exempel 11 följer också mönstret som visas i figur 2, den absolut största kategorin är strykningar men det förekommer även mycket ersättningar och minst tillägg. Med exempel 9 och 11 är det svårt att fokusera på något särskilt fel, utan det är antalet fel som gör dem så allvarliga. I båda exemplen är budskapet förvrängt eller flummigt, men i det finlandssvenska exemplet är alla ord begripliga. Däremot i det sverigesvenska exemplet har det strukits hela meningar och ett ord har ersatts med endast en bokstav *d*. En sådan här genomgång bevisar tydligt hur WER inte är ett entydigt sätt att betygsätta taligenkänningsverktyg. Avsaknaden av skiljetecken påverkar också läsbarheten och framhävs i längre exempel, men jag vill påminna om att det inte är i fokuset i min studie.

Jag har nu analyserat exemplen med lägsta och högsta felprocent och gett en översikt över WER-felkategorierna. Låt oss nu titta närmare på hurdana fel egentligen ingår i dessa felkategorier, hur stora kategorierna är och om det finns några gemensamma fel mellan språkvarianterna.

7.1 Ersättningar

Vi börjar med kategorin ersättningar. Ersättningar är alltså fel där verktyget har tagit med rätt antal ord, men orden har inte tolkats korrekt. Det kan handla om att ord böjs fel eller ord som har tolkats helt fel. I det finlandssvenska materialet är denna kategori störst med exakt sju procent och i det sverigesvenska materialet näst största med strax under tre procent. Eftersom många typer av fel faller inom denna kategori är det en mångsidig och bred kategori i båda språkvarianterna.

Den vanligaste frasen som ofta misstolkas i det finlandssvenska materialet var talarnas gemensamma inledningsfras *värderade talman*, *ärade talman* eller *herr talman*. Ordet *värderade* tolkades bl. a. som *varierade*, *vardera detta*, *var deras*, *vadderade*, *för delade*, och bara *de*. *Talman* tolkades ofta som *tar man*, *gör man*, *där man*, *dar man*, *där har man*, *dahlman* och *Danmark*. Således blev frasen *värderade talman* t.ex. *vardera detta Ahlman*, *vad deras talar man*, *det är talman*, *var deras tar man*, *detta har man*, *vardera det tar man*, *var deras tar man* och till och med *jag dödar ni*. Frasen *värderade herr talman* blev *det är vad*

det här talmannen. *Ärade talman* däremot tolkades som *är detta gör man*, *är detta allman* och även *enligt Ann-Marie*.

De allvarligaste felen kan tänkas vara tolkningar som ändrar innebörden av meddelandet eller gör det omöjligt att förstå. Det fanns väldigt många sådana fel, men det förekom mest med prepositionerna, pronomen och andra småord. *På* hade tolkats som *en* eller *att*, *till* tolkats som *i* eller *att*, *i* hade tolkats som *är* flera gånger och *från* tolkades som *ifrån*. *De*, *det* och *den* tolkades som *de*, *i*, *det*, *den* eller *denna*. Det överlägset vanligaste fallet var att *de* tolkades som *det*. Vidare tolkades bland annat *jag* som *att*, *vi* som *vid*, *så* hade blivit antingen *har*, *som* eller *och*, *en* tolkades som *är* och *är* antingen som *en*, *hur*, *de* eller *i*, *och* tolkades som *som*, *av*, *har* eller *att* och *som* tolkades som *att*. Det förekom också många enskilda fall, t.ex. ordet *ägo* hade blivit *ego*, frasen *kanske ett hade* hade blivit *kan tjäna* och *export av el* hade tolkats som *exporta väl*. *Läkarledamot* hade tolkats som *lek och ledamot* och vissa sammansatta ord som *mödrarrådgivningsbyrån* hade tolkats som två eller flera individuella ord, t.ex. *mödrars rådgivningsbyrån*. Sedan förekom det fall där frasen som *finansierar de* hade tolkats som ett ord *finansierade*. Det var också intressant att notera att *skulle* hade flera gånger tolkats som *ska* i det finlandssvenska materialet.

En mängd egennamn och förkortningar i det finlandssvenska materialet visade sig vara svåra för verktyget att tolka. *Ledamot Toivolas* hade tolkats som ett annat egennamn, *ledamot i Volvos* och *Jussi Aaltonen* som *Josefin Aaltonen*. Tågbanan *Bennäs-Jakobstad-Alholmen* blev tågbanan *Bennets Jakobstad anhållen men*. *EU* tolkades som *är ju* och *YE* och *EU-rapporten* blev *efter rapporten*. *UPM:s* blev *upp hems*, *TVO:s* blev *det vid Åhs*, *datasystem Aipa* blev *datasystem och Saipa* och *också Olkiluotos kärnkraftverk* blev *aldrig lova att kärnkraftverk*. Frasen *för Rosatom är det* blev *för oss att de är det* och ytterligare frasen *blir Rosatom den* blev *blir oss att om det*. I ett exempel tolkades Rosatom dock korrekt, *Rosatom som sitter i Kremls famn* blev *Rosatom som simmade hem hos familjen*. Även *Finland* feltolkades som *filmer* och *Finlands* som *finnas*.

Fel som är mindre allvarliga är tolkningar där suffixet har tappats bort. Ordet i sig förändras alltså inte, men däremot kan till exempel bestämdhet ändras. Det fanns ett riktigt stort antal sådana i materialet, här är några exempel, med det som verktyget har utelämnat inom parentes: *miljösamarbete(t)*, *betänkande(t)*, *förslag(et)*, *fokusera(r)*, *länderna(s)*, *stiftelse(r)*, *fungera(t)*, *fundera(r)*, *lite(t)*, *bli(r)*, *fatta(t)*, *beslut(et)*, *bekanta(t)*, *beroende(t)*, *samhälle(t)*, *uttala(t)* och *utvecklande(t)*. Det fanns också fall verktyget tillade ett suffix i slutet av ord,

dvs. ord som *öka*, *vår*, *ekonomi*, *skapa* och *problematisk* hade tolkats som *ökad*, *våra*, *ekonomin*, *skapar* och *problematiskt*. Och en del fall där ordets suffix ändrades vid verktygets tolkning, t.ex. *poliser* blev *polisen*, *priset* blev *priser*, *talas* blev *talar* och *råkar* blev *råkade*. Dessa räknas också som ersättningar (*miljösamarbetet* ersätts av *miljösamarbete*), även om jag talar om att suffixet har strukits eller lagts till.

I det sverigesvenska materialet förekommer det mycket liknande feltolkningar även om kategorin är mindre. *Tack herr talman* eller *herr talman* är den vanligaste frasen att inleda talet med och frasen tolkades som *tackar man*, *tucker man*, *tacka talman* och *här tar man*. *Ja, herr talman* tolkades som *jag har talman* och *herr* en gång endast som *är*. Av dessa förekom *här tar man* flera gånger och de övriga exemplen egentligen bara en gång.

De allvarligare felen i det sverigesvenska materialet bestod också delvis av småord. *De*, *det* och *den* hade tolkats som *man*, *som*, *det*, *dem*, *en*, *det* och *de*. Helt i motsats till det finlandssvenska materialet var det vanligaste felet att *det* tolkades som *de*. *Ja* tolkades som *jag*, *vi* tolkades antingen som *man*, *vid* eller *vill*, *så* tolkades som *för*, *är* antingen som *med*, *om* eller *gör*, *ett* som *så*, *er* som *min* eller *i*, och *som* tolkades som *om*. Ord som *var*, *vara*, *varit* och *vad* tolkades alla ofta som *vad* eller *var* och någon gång som *och*. *Peng* tolkades som *poäng* och *att regera* som *attrahera*. Fras som *klä dem på morgonen* blev *kläderna på morgonen*. En icke-standardiserad sammansättning *hästköttslasagnen* tolkades helt enkelt som *hästkött i lasagnen* och fras *i vår paj eller i vår lasagne* blev *i vår pipeline i eller i vår lasagne*. Sedan förekom det också fall där ord som *till för* hade tolkats som *tillför*.

Feltolkningar förekom också i fråga om egennamn och förkortningar. Namn som *Faribah* och *Maj-Britt* hade tolkats som *fria* och *MAI Britt*. *Jimmie Åkesson* däremot hade tolkats som *ger mig också* och ytterligare *jag och Jimmie Åkesson lever i olika* som *jag ger mig också leva i olika*. Frasen *Peter Erikssons dagar* blev *Peter Eriksson hans dagar*. *Facebook* hade tolkats som *friskolor*, *OECD-länder* som *oeniga länder* och *RUT-bidrag* som *RUT-avdrag*.

Det fanns också likheter i tolkningar av suffix i det sverigesvenska materialet. I följande exempel har suffixet utelämnats från tolkningen *stoltsera(r)*, *lyfte(r)*, *omvärdera(r)*, *fungera(r)*, *markera(r)*, *kom(mer)*, *skapa(r)*, *arbeta(r)*, *verka(r), ta(r)*, *se(r)*, *börja(r)*, *stoppa(r)*, *finansiera(r)*, *underlätta(r)*, *rösta(r)*, *fixa(r)*, *misshandla(r)*, *begränsa(r)*, *handla(r)*, *jobba(r)*, *tala(r)*, *förkorta(s)*, *förändringar(na)*. I följande ord har suffio antingen lagts till eller ändrats: *använd* blev *använder*, *låt* blev *låter*, *hus* blev *huset*, *förbjud* blev *förbjuda*,

skolstart blev *skolstarten*, *välkomna* blev *välkomnar*, *vinstintresse* blev *vinstintresset* och sedan *samhällen* blev *samhället*, *fördelar* blev *fördelade*, *förändringen* blev *förändringar* och *målen* blev *målet.*

Det som gör denna kategori den största eller näst största är säkerligen det faktum att kategorin faktiskt täcker många olika typer av fel, vissa mer allvarliga och andra mindre allvarliga. En annan orsak till att denna kategori är så stor är att de flesta av felen inträffade bara en gång. Kategorin innehåller därför många exempel, särskilt ur finlandssvenskt materialet, men exemplen i sig förekommer få gånger. Ersättningar ger också ofta upphov till andra fel, t.ex. då frasen *ni är bäst* tolkas som *nio bäst,* ordet *nio* ersätter ordet *ni* och då identifieras ordet *är* egentligen som en strykning. Av denna anledning innehåller de exempel som presenteras i denna kategori ofta även andra typer av fel. För tydlighetens skull har jag dock inte färgkodat de övriga felen i detta kapitel. Denna kategori innehåller alltså fel som är gemensamma för båda språkvarianterna, men kategorin är helt klart mindre i det sverigesvenska materialet. Bland de allvarligare felen i det sverigesvenska materialet saknades prepositioner helt och fel som handlade om suffix var egentligen sådana där presens hade tolkats som grundform. I det finlandssvenska materialet åtföljdes alla dessa fel även av en mängd andra typer av fel, vilket bevisar skillnaden i kategoristorlek mellan språkvarianterna.

7.2 Strykningar

Nästa kategori är strykningar. Alltså fel där verktyget har tappat bort ord och således saknas det ord i taligenkänningens tolkning. Denna kategori är den minsta kategorin i det finlandssvenska materialet med cirka en och en halv procent och den största felkategorin i det sverigesvenska materialet med drygt tre procent. Det är således tydligt att kategorierna är olika stora i de två språkvarianterna. Det som finlandssvenskan och sverigesvenskan dock har gemensamt i den här kategorin är att den innehåller ett brett spektrum av ord, från pronomen till substantiv och verb.

Ord som har strukits i det finlandssvenska materialet är *är*, *jag*, *det*, *de*, *så*, *i*, *talman*, *har*, *värderade*, *ju*, *en*, *på*, *så*, *år*, *att*, *skulle*, *som*, *ett*, *av*, *kan*, *och*, *herr*, *här*, *vi*, *bör*, *också*, *tack*, *för*, *men*, *kommer*, *till*, *gällt*, *egen*, *oss*, *vill* och *sades*. Ordet *är* strukits oftast och de två följande orden *jag* och *det* utelämnas också ofta, men de återstående orden förekommer bara en eller två gånger.

I det sverigesvenska materialet motsvarande ord är *det*, *så*, *ju*, *är*, *här*, *då*, *och*, *jag*, *vi*, *att*, *från*, *en*, *jo*, *nu*, *faktiskt*, *du*, *tack*, *tycker*, *er*, *om*, *den*, *kommer*, *talman*, *ja*, *a-kassan*, *ges*, *man*, *vad*, *mycket*, *dem*, *hem*, *liksom*, *ni*, *den*, *för*, *någonting*, *egentligen*, *med*, *helt*, *herr*, *var*, *Socialdemokraterna*, *folket*, *förr*, *eller*, *debatten*, *inför*, *väl*, *i*, *påstå*, *på*, *i dag, har* och *de*. På samma sätt är orden i den ordning i vilken strykningarna förekom mest. Det mest raderade ordet det i sverigesvenskan förekom dock flera tiotals gånger oftare än det mest raderade ordet i finlandssvenskan. I allmänhet var det i sverigesvenskan fler ord som ströks mycket ofta och färre som bara förekom en gång.

Även om volymerna inom denna kategori varierar är det mycket som är gemensamt mellan de två språkvarianterna. Bland de mest borttagna orden ligger samma ord, *det* och *är*. Detsamma gäller också för det faktum att kategorin inte är begränsad till någon specifik ordklass i ingen av språkvarianterna. Antalet exempel visar att denna kategori är större i sverigesvenskan och att verktyget har strukit överraskande långa ord ur det sverigesvenska materialet medan fel ur finlandssvenskt material är koncentrerade till kortade ord.

7.3 Tillägg

Sista felkategorin är tillägg. Tillägg är feltolkningar som läggs till av verktyget, dvs. ord som inte har sagts av talaren och finns inte i transkriptionen. I det finlandssvenska materialet är denna kategori näst största med mer än två och en halv procent. I det sverigesvenska materialet är denna kategori betydligt mindre än de två andra, bara drygt en procent av felen tillhör hit. I båda språkvarianterna omfattar denna kategori mestadels en mängd småord, dvs. prepositioner, pronomen och konjunktioner. Det vanligaste ordet som lades till var detsamma i båda varianterna och förekom betydligt oftare än något annat tillägg, nämligen ordet *och*.

Andra ord som lades till i det finlandssvenska materialet var *att*, *av*, *ett*, och *det*. Av dessa lades *att* till nästan lika ofta som det vanligaste ordet *och*. De övriga orden lades till mycket mer sällan, men sinsemellan ungefär lika ofta. Vid ett tillfälle i det finlandssvenska materialet lades till även siffran *1*. Denna kategori har därför många likheter med den tidigare kategorin strykningar i finlandssvenskan.

Utöver ordet *och* förekom följande tillägg i det sverigesvenska materialet; *att*, *är*, *det*, *den*, *de*, *jag*, *en*, *vid*, *i*, *för*, *på*, *av*, *om*, *när*, *tror*, *annat* och *har*. Av dessa förekom de två första orden, *att* och *är*, i ett betydande antal och resten i genomsnitt endast en eller två gånger.

Det vanligaste och näst vanligaste orden som lades till av taligenkänningen var alltså detsamma i finlandssvenskan och sverigesvenskan trots att storleken på kategorin skiljer sig åt mellan de två språkvarianterna. I allmänhet bestod denna kategori av mycket liknande ord och ordklasser för båda språkvarianterna. Eftersom denna kategori är mycket större för finlandssvenskan än för sverigesvenskan kan det tyckas märkligt att det finns fler exempel här för sverigesvenskan. Detta kan förklaras med att i finlandssvenskan var dessa fel ofta relaterade till eller orsakade av ett annat fel, t.ex. ett ersatt ord, och därför behandlas en stor del av tilläggen i det finlandssvenska materialet i samband med ersättningar i kapitel 7.1 eller i kapitel 8. Däremot i det sverigesvenska materialet var dessa fel ofta inte relaterade till andra fel och därför framhävs felen i det sverigesvenska materialet i detta kapitel.

I det här kapitel har jag illustrerat hur de tre felkategorierna ser ut i de två språkvarianterna. Jag har analyserat två exempel per språkvariant samt sammanfattat varje felkategori. I nästa kapitel fokuserar vi på vad som kan ha orsakat dessa fel.

8 Möjliga orsaker till taligenkänningsverktygets feltokningar

I detta kapitel diskuterar jag resultaten ur ett kvalitativt perspektiv. Jag presenterar möjliga orsaker till de fel som beskrevs och kategoriserades i föregående kapitel 7. Ytterligare försöker jag ta reda på om orsaker bakom felen är desamma i sverigesvenska och finlandssvenska.

Den första kategorin som behandlades i föregående kapitel var ersättningar, som egentligen bestod av flera mindre kategorier. Inledande ord i det finlandssvenska materialet innehöll orden *värderade*, *ärade*, *herr* och *talman*. I det sverigesvenska materialet var det betydligt vanligare att inleda talet helt enkelt med *tack, herr talman*. I båda språkvarianterna förekom det feltolkningar i dessa inledningsfraser, men mycket mera i det finlandssvenska materialet. Den enklaste förklaringen till dessa feltolkningar kan vara att taligenkänningen känner inte igen dessa ord. Fraser som *ärade talman* eller *värderade talman* låter väldigt gammalmodiga om man tar dem ur det politiska sammanhanget. Här spelar taligenkänningsverktygets språkmodell en viktig roll (Papastratis 2021). Valet av en språkmodell för taligenkänningen styr t.ex. om verktyget ska fokusera på ett visst ämnesområde. Eftersom det fanns bara en språkmodell för finlandssvenskan i verktyget i fråga valde jag på basis av detta för sverigesvenskan en språkmodell som liknande den för finlandssvenskan, således var språkmodellen för båda språkvarianterna ”generell”. Med tanke på materialets begränsade ämne och det faktum att dessa inledande fraser är mycket typiska för ämnet, kunde dessa feltolkningar ha förhindrats genom att välja en annan språkmodell. På så sätt skulle kanske till och med en mindre, men väl sammanställd språkmodell ha kunnat minska vissa typer av fel jämfört med en större, mer generell språkmodell. Detta kan dock bara spekuleras.

Ibland lyckades verktyget tolka dessa inledningsfraser korrekt, detta illustreras t.ex. i exempel 11 på sidan 40 f. Exempel 11 är ur sverigesvenskt material och som det har konstaterats förekom fler av dessa fel i det finlandssvenska materialet. Detta kan bero på att den vanligaste inledningsfrasen *(tack) herr talman* i det sverigesvenska materialet är inte lika konstig när det tas ur sammanhanget. Här kan vi se att finlandssvenskan har hållit fast vid de gammalmodiga inledningsfraserna (Reuter 2015a:24 ff.). De tillfällen då de inledande fraserna tolkades korrekt i det finlandssvenska materialet kan förklaras av med god artikulation (Forsberg 2003:7). God artikulation hänger ofta ihop med talarens ålder. Bland de finlandssvenska talarna fanns det flera äldre personer vars något mindre välartikulerade tal påverkade

taligenkänningens framgång. Den tydliga skillnaden var att de yngre talarna talade tydligare vilket lesse till färre feltolkningar.

En annan förklarande faktor vid tolkningen av just dessa inledande fraser kan vara omständigheterna vid tillfället, som gäller för båda språkvarianterna. Dessa parlamentariska evenemang har alltså flera talare, så den person som har fått ordet reser sig upp och går fram till publiken där hen kan tala in en mikrofon. Det hände ofta att de finlandssvenska talarna började tala redan då de var på väg fram till publiken och inte stod framför mikrofonen, så att de första orden i tal, dvs. de inledande fraserna, blev mycket svåra att höra. Detta gör det också svårare för taligenkänningsverktyget att tolka dessa ord korrekt (Arora och Singh 2012:37). De sverigesvenska talarna var mer benägna att vänta till de stod framför mikrofonen så det blev färre feltolkningar av de första orden. Å andra sidan innebär omständigheterna att talarna har begränsad tid. Svenskarna kan ha försökt kompensera för den begränsade tiden genom att tala snabbare, medan finländarna kan ha försökt spara tid genom att börja sitt tal redan på väg till publiken.

Dessa orsaker bidrar helt klart till en lång rad feltolkningar, men det kan finnas fler orsaker. Fallet där *export av el* tolkades som *exporta väl* handlar till exempel om det som Engstrand (2016:18 f.) beskrev i sitt experiment, att språkljud glider från det ena till det andra. Särkilt om man talar snabbt är det mycket svårt tolka var ordgränserna ligger i tal. Säger man snabbt *exportavel* kan det alltså vara omöjligt för lyssnaren att avgöra om vokalen /a/ är en del av ordet *export* eller inte och om konsonanten /v/ börjar eller slutar ett ord. Feltolkningen där *finansierar de* tolkades som *finansierade* beror förmodligen på att det sista r-ljudet i ordet *finansierar* kan ha fallit bort i talet och då är det återigen svårt att tolka om det följande ordet *de* är ett eget ord eller ett suffix på det föregående ordet. Det är inte förvånande att ordet *hästköttslasagnen* har misstolkats, eftersom det inte finns något sådant, men det är förvånande att verktyget har tolkat innebörden av ordet perfekt, *hästkött i lasagnen*. Den lösning som verktyget tillhandahåller är därför faktiskt genial, men vad gäller studien behövs det ändå markeras som en feltolkning.

Vissa fel beror också på skillnader i uttalet. När förkortningen *TVO:s* i det finlandssvenska materialet tolkades som *det vid Åhs* är orsaken till detta att finlandssvenskans [t] är oaspirerad och tolkas därför som något med [d] (Reuter 2015a:24 ff.). Det är också möjligt att förkortningen inte är bekant för verktyget. Det finns fler exempel på de oaspirerade klusiler som är karakteristiska för finlandssvenskan, också i inledningsfraserna där *talman* har tolkats

bl. a. som *dahlman* eller till och med *Danmark*. I det sverigesvenska materialet har *talman* ofta feltolkat som *tar man*, dvs. feltolkningen beror inte på aspirationen. När det gäller vokalkvaliteten var det intressant att i det finlandssvenska materialet var *EU-rapporten* tolkats som *efter rapporten*. Detta kan möjligen tyda på att talaren har uttalat *EU* mer som /ev/ än /eu/, vilket skiljer sig något från listan över uttalsskillnader (Reuter 2015a:21–24). Liknande feltolkningar relaterade till uttalsskillnader förekom inte i någon större utsträckning i det sverigesvenska materialet. Även om de språkmodeller som användes i studien var specifika för språkvarianterna, är det kanske ändå möjligt att tänka sig att skillnaderna i uttalet mellan finlandssvenska och sverigesvenska ledde till vissa feltolkningar i det finlandssvenska materialet.

Feltolkningar av egennamn och vissa förkortningar, till exempel *Åkesson* och *u-land*, kan antas bero på att de inte är bekanta för verktyget, dvs. verktyget inte kan känna igen och tolka dessa ord korrekt. Några egennamn identifierades dock korrekt av verktyget, åtminstone med den finlandssvenska språkmodellen, eftersom *Toivola* och *Aipa* tolkades som *Volvo* och *Saipa*. Det är intressant att egennamnet *Facebook* i det sverigesvenska materialet identifierades inte av verktyget, utan tolkades som *friskolor*, trots att *Facebook* kan tänkas vara en plattform som alla kände till år 2014.

Fel där suffixet har utelämnats i tolkningen kan bero på flera olika orsaker. Feltolkningar kan vara relaterade till andra fel, som i exemplet med *finansierar de*. Fel kan också bero på talhastigheten och hur orden assimileras ihop (Engstrand 2016:15) eller helt enkelt på att talaren inte har uttalat slutet av orden (Engstrand 2016:22 och 2012:12 f.). Av antalet exempel kan vi se att sådana feltolkningar förekommer mer i det finlandssvenska materialet, vilket ligger i linje med Reuters observation att ordändelser uttalas hörbart oftare i sverigesvenskan än i finlandssvenskan (Reuter 2015a:31).

Denna kategori omfattar ett brett spektrum av fel och därför är orsakerna till dessa fel mångsidiga och olika. Valet av en språkmodell kan ha orsakat svårigheter med taligenkänningen, men även många andra orsaker kunde identifieras, t.ex. skillnader i uttal, artikulation, talhastigheten och talarnas ålder.

Därefter tittat vi på felkategorin strykningar. De vanligaste orden som strukits i båda språkvarianterna var *är* och *det*. Anledning till detta är att dessa ord är så korta och små att de ofta smälter samman med ett annat ord i tal (Engstrand 2016:18 f.). Materialet innehöll till

exempel fallet där *syftet är* hade tolkats som *syftar*. Ordet *är* har förmodligen uttalats bara som ett snabbt r-ljud och därför har taligenkänningen tolkat att r-ljudet egentligen hör till ordet före det och således har ordet *syftet* blivit *syftar* och order *är* stryks. På liknande sett förekom det ofta att frasen *har det* hade tolkats som *hade*. Här är det också tydligt hur orden *har* och *det* smälter samman i tal och som en följd av det stryks ordet *det*. Fallet i exempel 6, där ordet *kan* stryks och ordet *placeras* ersättas med ordet *kontrolleras*, kanske inte är lika tydligt men även där kan vi se att *kan placeras* innehåller många samma konsonanter som ordet *kontrolleras* och därmed är taligenkänningsverktygets feltolkning ändå rimlig.

De flesta av de strukna orden som förekom ofta var alltså småord som naturligt förknippas med andra ord i tal, men i det sverigesvenska materialet stryks även långa ord eller hela meningar som illustreras i exempel 11. I exempel 11 har meningar *jag lovar* och *ni får vara med* utelämnats helt av verktyget. Från materialet ser vi att också enskilda längre ord som *någonting*, *egentligen* och *Socialdemokraterna* stryks. Ibland, särskilt då flera ord i följd har strukits, förekom det bakgrundsbuller i ljudklippen, vilket säkert har påverkat verktygets tolkning (Arora och Singh 2012:37 och O'Shaughnessy 2008:2966). Men eftersom man i allmänhet lyssnar på talaren i politik är det mer sannolikt att orsakerna till strykningar är talhastighet och att orden förkortas mycket. Som framgår av Engstrands exempel på sidan 18 kan även ett långt ord som *faktiskt* reduceras till så få som tre språkljud (Engstrand 2016:63 och 2012:12, 16). Talhastighet och förkortning av ord går därför ofta hand i hand.

I exempel 11, som fick den högsta felprocenten ur sverigesvenskt material, säger talaren mer än 200 ord i minut. Som jämförelse kan nämnas att i exempel 9, som fick den högsta felprocenten ur finlandssvenskt material, ligger talhastigheten på 112 ord i minut. Generellt sett talar svenskarna snabbare med en genomsnittlig talhastighet på 162 ord i minut. Finlandssvenskarnas genomsnittliga talhastighet var betydligt lägre, 124 ord i minut. När orden blandas ihop i snabbt tal är det utmanande för verktyget att tolka tal, men inte nödvändigtvis för den mänskliga hjärnan (Engstrand 2016:21 ff). Eftersom det finns så stora skillnader i talhastigheten kan det vara orsaken till att kategorierna är så olika stora mellan språkvarianterna. I det sverigesvenska materialet är dessa fel mer uttalade eftersom svenskarna talar mycket snabbare och eftersom finlandssvenskarna talar mycket lugnare är dessa fel mindre frekventa.

Som vi kan se i exempel 1 och 2 på sidan 29 består det sverigesvenska materialet av mycket mer personlig interaktion. Karaktär i det sverigesvenska materialets är i allmänhet mer

diskursiv än i det finlandssvenska materialet. Detta beror på att det sverigesvenska materialet bygger på partiledardebatt, alltså talet är mer konversationellt. Detta leder till mer personliga samtal och förhöjda känslor hos sverigesvenska talare, vilket bidrar till en snabbare talhastighet. Det finlandssvenska materialet däremot samlades in under ett halvt år, så talturen tar inte alltid vid där det föregående slutade. Bland finlandssvenskarna är debatten generellt sett lugn och allt kommenteras inte, särkilt inte på svenska, eftersom de flesta i publiken inte talar eller förstår svenska. Detta innebär att de finlandssvenska talturen är åtskilda från varandra och att det inte uppstår något känslomässigt tumult. Lugnare, mer okänsligt tal kan vara orsaken till vissa strykningar, eftersom verktyget kanske inte kan fånga upp allt från tystare tal, men å andra sidan hjälper en tyst talfrekvens verktyget att känna igen fler ord bättre (Forsberg 2003:7). Som nämnts har talarna vid parlamentssessioner en begränsad tid att tala, vilket kan leda till att de rusar iväg och talar snabbt. Detta gäller dock för båda språkvarianterna.

Inom denna kategori är orsakerna till fel alltså desamma för båda språkvarianterna, dvs. talhastighet, förkortning av ord och eventuell bakgrundsbrus. I sverigesvenskan är dock dessa orsaker och därmed felen mer uttalade.

Sista felkategori är tillägg och det mest tillagda ord var gemensamt för båda språkvarianterna och det var odet *och*. Övriga tillägg var i allmänhet andra småord och de finlandssvenska tilläggen var ofta relaterade till en annan felkategori, medan de sverigesvenska tilläggen inte var det. Detta bidrar säkert till att detta var den näst största kategorin i det finlandssvenska materialet, eftersom ersättningar också producerade fel i denna kategori. Eftersom detta inte var fallet i det sverigesvenska materialet var denna kategori överlägset minst där.

Tillägg som har helt korrekt tolkat text på båda sidorna, dvs. där det tillagda ordet inte kan kopplas till något annat fel, är svåra att analysera. Det kan vara svårt att tyda varifrån det tillagda ordet kommer när det inte finns några andra fel i närheten. Till exempel en feltolkning där *tingsrätterna* har blivit *tingsrätten att* kan anses vara rimlig eftersom man kan se hur felet har uppstått. Talaren har sannolikt inte betonat slutet av ordet *tingsrätterna* (Reuter 2015a:29 ff.), så det enda ljudet som framhävs i slutet kan har varit vokalen /a/ och således har verktyget tolkats vokalljudet som ett eget ord *att*, vilket blir ett tillägg. Ordet *tingsrätten* däremot identifieras som en ersättning av ordet *tingsrätterna*, dvs. även i detta fall ledde ersättning av ett ord till att ett ord tilläggs.

Det vanligaste tillägget *och* gjordes på ställen där det annars skulle ha varit en punkt, dvs. mellan meningarna. I många fall utelämnades punkter och andra skiljetecken för att talaren inte hade lagt till dem i sitt tal, eftersom det är den princip som verktyget i fråga fungerar enligt. Början och slut på meningar kan i tal markeras med en liten paus och ett andetag. Detta kan vara en naturlig förklaring till att en så liten inandning ofta tolkas som ordet *och* av taligenkänningen. Talhastigheten kan återigen förklara varför denna typ av tillägg var vanligare i finlandssvenskt material. Eftersom finlandssvenskarna talar långsammare kan de ha haft mer tid att pausa mellan meningarna, dvs. taligenkänningen ville eventuellt oftare lägga till ett ord i dessa utrymmen. De sverigesvenska talarna har förmodligen inte tagit så långa pauser så verktyget har inte hunnit göra några egna tillägg. Den största kategorin i det sverigesvenska materialet var strykningar, vilket möjligen stöder detta fenomen, dvs. att talet verkligen har varit så snabbt att mycket har utelämnats ur tolkningen snarare än att något extra har lagts till.

Det finns en feltolkning i denna kategori som handlar åtminstone delvis om uttal. I det finlandssvenska materialet fanns ett ställe där *Fennovoima* hade tolkats som *vän och*. Språkljuden [f] och [v] är båda labiodentala frikativa där [f] är tonlösa och [v] är tonande, dvs. ljuden ligger mycket nära varandra. Konsonanten /f/ kan ibland i tal framträda som den tonande labiodentala frikativa [v], vilket är naturligt i finskan och kan ha påverkat en finlandsvensk talares tal. Detta exempel har hamnat i kategorin tillägg, eftersom ordet *och* har lagts till, men det finns också en ersättning av ordet *Fennovoima* med ordet *vän*.

Som redan konstaterats var tillägg i det finlandssvenska materialet ofta relaterade till andra felkategorier, som i exemplet med *Fennovoima* ovan. Så var inte fallet i det sverigesvenska materialet. Andra tillägg som förekom, särskilt i det sverigesvenska materialet, var småord, som *att*, *är* och *det*. Jag misstänker att orsakerna till att det finns tillägg i det sverigesvenska materialet som inte kan förknippas till andra felkategorier delvis är desamma som för strykningarna. På samma sätt som småord kan vara lätta att stryka eftersom de lätt smälter in i de omgivande orden, är det också lätt att lägga till småord i tolkningar, eftersom biljud eller andningsljud i snabbt tal kan redan bilda något som verktyget kan tolka som prepositioner, t.ex. *till* och *för*.

I det finlandssvenska materialet utgjordes tilläggen alltså mest av andra felkategorier, framför allt av ersättningar. I det sverigesvenska materialet fanns det få tillägg, men de var mer svårtolkade. Talhastighet och ordassimilation var dock troliga orsaker till dessa fel.

I det här kapitel har jag undersökt möjliga orsaker till de feltolkningar som uppstod vid taligenkänning. Jag har försökt ta reda på om det finns några gemensamma orsaker till de fel som är vanliga i språkvarianterna. Kategorin ersättningar är den största kategorin i det finlandssvenska materialet och den näst största i det sverigesvenska materialet. Vanliga orsaker till feltolkningar var bland annat talarnas ålder, artikulation, assimilation samt valet av en språkmodell. I det finlandssvenska materialet förekom också en del feltolkningar som möjligen berodde på uttalet. Eftersom kategorin var betydligt större i det finlandssvenska materialet förklarar detta sannolikt varför det finlandssvenska materialet lyfter fram en särskild orsak till vissa av felen som inte förekommer i det sverigesvenska materialet. Kategorin strykningar är den minsta kategorin i det finlandssvenska materialet och den största kategorin i det sverigesvenska materialet. De vanligaste strukna orden i båda språkvarianterna var *är* och *det*, vilket naturligtvis förklaras av att dessa ord är högfrekventa i svenskan. I båda språkvarianterna förklaras strykningarna av att de ofta smälter samman med andra ord. I det sverigesvenska materialet ströks det även överraskande långa ord, vilket sannolikt förklaras av att sverigesvenskar talar mycket snabbare än finlandssvenskar. Kategorin tillägg är den näst största kategorin i det finlandssvenska materialet och den minsta i det sverigesvenska materialet. De mest tillagda orden var gemensamma för båda språkvarianterna och dessa tillägg har troligen lagts till i stället för punkter av taligenkänningen. Att kategorin är så stor i det finlandssvenska materialet beror till stor del på att dessa fel är starkt förknippade med ersättningar. Felen i det sverigesvenska materialet var mer svårtolkade, men talhastighet och assimilation är troliga orsaker.

Om vi ser till den övergripande bilden i rent numeriska termer, är den genomsnittliga WER-felprocent för det finlandssvenska materialet 0,14 % och motsvarande felprocent för det sverigesvenska materialet är 0,07 %. Resultatet är förvånande, eftersom den höga talhastigheten som betonas i det sverigesvenska materialet verkade orsaka en rad olika problem med taligenkänning. Det verkar dock som att den hörbara rösten, tydlig artikulation och yngre talare i det sverigesvenska materialet hade större inverkan i resultatet. Trots en så stor skillnad i genomsnittlig WER presterade båda språkvarianterna ändå bra enligt Kakkar (2023).

9 Sammanfattande diskussion

Huvudsyften med denna studie var att undersöka om det finns signifikanta skillnader i taligenkänning mellan finlandssvenska och sverigesvenska, som är två nationella varieteter av svenska. Syftet var också att undersöka orsaker till eventuella skillnader i taligenkänningen, om de beror på uttalsskillnader mellan språkvarianterna eller på andra faktorer och slutligen, hur kan taligenkänningsverktyg förbättras. Material som används i studien är riksdagsplenum och partiledardebatt från 2014. Studien krävde både kvantitativa och kvalitativa metoder. Jag använde WER (*word error rate*) för att utvärdera taligenkänningsverktyget och en komparativ metod för att identifiera och analysera taligenkänningens feltolkningar.

Det finns en stor skillnad i den genomsnittliga felprocenten för de två språkvarianterna. Den genomsnittliga felprocenten för finlandssvenskan är 0,14 % medan felprocenten för sverigesvenskan är mycket lägre, 0,07 %. Det kan därför konstateras att taligenkänningsresultaten för sverigesvenskan var dubbelt så korrekta och exakta som för finlandssvenskan. Orsaker till taligenkänningens feltolkningar var delvis desamma för de två språkvarianterna, till exempel talarnas ålder och artikulation, vilket försämrade prestandan för finlandssvenska och förbättrade prestandan för sverigesvenska. Förkortning och assimilering av ord förekommer också i båda varianterna. Utöver detta framkom några orsaker som var specifika för en av språkvarianten, för det finlandssvenska materialet verkade det vara typiskt att uttalet hade en viss inverkan på feltolkningar och för det sverigesvenska materialet en svindlande talhastighet.

Även om WER inte är ett helt svartvitt sätt att bedöma kvaliteten på taligenkänning, är skillnaden mellan språkvarianterna så stor att det inte kan ignoreras. I det sverigesvenska materialet låg fokus alltså på den höga talhastigheten. I det finlandssvenska materialet talade talarna i allmänhet lungt och långsammare. Det verkar därför som om talhastigheten inte hade någon betydande effekt på resultaten. Talarnas ålder, deras artikulation och tydlighet i talet skapade dock skillnader mellan språkvarianterna. Detta bevisar att de välkända utmaningarna vid taligenkänning fortfarande skapar problem (Forsberg 2003:7 samt Arora och Singh 2012:37). Vissa feltolkningar på grund av uttalsskilnnader förekom också (Reuter 2015b:21–31), men i minskande antal.

Det är något oklart om man kan säga att de huvudsakliga skillnaderna mellan språkvarianterna inte beror på särdrag som är specifika för pluricentriska språk (Clyne 1992b:passim). I det

finlandssvenska materialet förekom en del feltolkningar som verkar bero på uttalet. Liknande fel förekom inte i det sverigesvenska materialet, vilket leder till slutsatsen att det var just det finlandssvenska uttalet som gav upphov till feltolkningarna. I studien användes dock en egen språkmodell för var sin språkvariant, dvs. man skulle kunna anta att en språkmodell som riktar sig till finlandssvenskan skulle ignorera fenomen som är typiska för finlandssvenskan och inte skulle bidra till feltolkningar. Detta kan möjligen bero på att det har funnits mer material och träningsdata för sverigesvenskan, vilket gör att taligenkänningen fungerar bättre för sverigesvenskan.

Mängden träningsdata kan också förklara varför taligenkänningen på finlandssvenskan var så mycket sämre än på sverigesvenskan. De finlandssvenska talarna talade mer skriftspråksnära medan de sverigesvenska talarna talade mer spontant och naturligt. Engstrand (2012:9) och Douglas O'Shaughnessy (2008:2966) menar att detta borde leda till motsatta forskningsresultat. Å andra sidan om undersökningsresultaten beror på mängden träningsdata skulle det vara i linje med min första känsla, dvs. att data är svårare att samla in om det finns färre talare av ett språk eller en språkvariant. Detta fenomen återspeglades också i urvalet och insamlingen av forskningsmaterialet.

Eftersom materialet i den här studien bestod av förinspelat tal och talare som därför inte kunde ta hänsyn till taligenkänning, skulle det vara viktigt att upprepa studien med informanter. Med hjälp av informanter kan uppmärksamhet ägnas åt deras ålder, artikulation och talhastighet och vilka av dessa eller vilka andra orsaker orsakar de största svårigheterna i taligenkänningen, dvs. om mängden träningsdata väger tyngre än till exempel artikulation. En sådan forskning skulle alltså kunna ge svar på frågan om det enda sättet att förbättra taligenkänning är att samla in mer och bättre data. Ytterligare skulle det vara intressant att se hur resultaten påverkas om talaren kan anpassa sitt tal efter taligenkänningsverktyg. Det är troligt att mängden träningsdata och valet av en lämplig språkmodell kommer att ha en större inverkan på resultaten. Jag tror alltså att nyckeln till bättre taligenkänning är att insamla mer data och bättre data. Detta är något att tänka på, eftersom t.ex. äldre människor är en av de grupperna vars taligenkänning är svårt på grund av nedsatt artikulation (Forsberg 2003:7 samt Arora och Singh 2012:37). Samtidigt är de också en grupp som kan ha stor nytta av talkommandon när deras finmotorik försämras (Tillgänglighetskrav.fi). Det bör också noteras att en mer ändamålsenlig språkmodell i sig skulle kunna ge bättre resultat, liksom bättre ljudåtergivningsutrustning (Papastratis 2021).

Studien visade alltså en tydlig skillnad i hur verktyget behandlar finlandssvenskt och sverigesvenskt tal. I takt med att olika tekniska lösningar och hjälpmedel blir allt vanligare är detta något som måste tas med i beräkningen. När sådana här verktyg används i praktiken kan det inte vara så att det är väldigt stora skillnader i hur de fungerar på till exempel majoritets- och minoritetsspråk. För att skapa ett mer tillgängligt samhälle för alla måste man också ta hänsyn till olika minoritetsgrupper i samhället. Denna studie visar att det fortfarande finns ett behov av att hitta lösningar för datainsamling, så att t.ex. talare av de icke-dominerande språkvarianterna av världens pluricentriska språk kan dra nytta av funktionell taligenkänning i framtiden.

Litteratur

Aalto-yliopisto, Signaalinkäsittelyn ja akustiikan laitos. *Aallon puheentunnistuskorpus eduskunnan istuntojen ruotsinkielisistä puheenvuoroista 2015-2020* [korpus]. Kielipankki. Saatavilla https://urn.fi/urn;nbn;fi;lb-2022052004.

Arora, Shipra J. & Singh, Rishi Pal, 2012: Automatic speech recognition: a review. I: *International Journal of Computer Applications*, 60(9). S. 34–44.

Cem, Dilmegani, 2023: AutoML: In-depth guide to automated machine learning in 2024. I: *AIMultiple*. https://research.aimultiple.com/auto-ml/. Hämtad 6 april 2024.

Cem, Dilmegani, 2024: Artificial intelligence (AI): In-depth guide for 2024. I: *AIMultiple*. https://research.aimultiple.com/ai/. Hämtad 6 april 2024.

Clyne, Michael, 1992a: Pluricentric languages – introduction. I: Clyne, M. (red.): *Pluricentric Languages. Different Norms in Different Countries*. Berlin/New York: Mouton de Gruyter. S. 1–9.

Clyne, Michael, 1992b: Epilogue. I: Clyne, M. (red.): *Pluricentric Languages. Different Norms in Different Countries*. Berlin/New York: Mouton de Gruyter. S. 455–465.

Digin.nu. Tillgänglighetsdirektivet. https://digin.nu/krav-och-regler/tillganglighetsdirektivet/. Hämtad 6 april 2024.

Engstrand, Olle, 2004: *Fonetikens grunder*. Lund: Studentlitteratur.

Engstrand, Olle, 2012: *Hur låter svenskan, ejengklien?* Riga: Norstedts.

Engstrand, Olle, 2016: Kan du säga Schweiz? En bok om uttal på svenska och utländska. Stockholm: Morfem.

Europaparlamentets och rådets direktiv (EU) 2019/882 av den 17 april 2019 om tillgänglighetskrav för produkter och tjänster. EUT L 151/70, 7.6.2019, s. 70. https://eur-lex.europa.eu/legal-content/SV/TXT/HTML/?uri=CELEX:32019L0882&from=SV#d1e40-70-1.

Forsberg, Markus, 2003: Why is Speech Recognition Difficult? Chalmers University of Technology. S. 1–10.

Gacek, Ann, 2020: Varför tillgänglighet är viktigt för din affär. Nå fler i din publik. I: *Kan.se.* https://kan.se/nyheter/varfor-tillganglighet-ar-viktigt-for-din-affar/. Hämtad 6 april 2024.

Gülbahar, Karatas, 2024: Speech Recognition: Everything you need to know in 2024. I: *AIMultiple*. https://research.aimultiple.com/speech-recognition/. Hämtad 6 april 2024.

International Working group on Non-dominant Varietes of Pluricentric Languages. https://pluricentriclanguages.org. Hämtad 17 september 2023.

Ivars, Ann-Marie, 1991: Finlandssvenskans ställning i förhållande till det svenska rikssspråket. I: *Språkbruk 4*. S. 3–6.

Juang, Biing-Hwang & Rabiner, Lawrence R., 2005: Automatic Speech Recognition – a brief history of the technology development. Santa Barbara: Atlanta Rutgers University and the University of California. S. 1–24.

Jurafsky, Dan & Martin, James H., 2023: *Speech and language processing*. 3:e upplaga under arbete. https://web.stanford.edu/~jurafsky/slp3/. Hämtad 16 april 2023.

Kakkar, Kushal S., 2023: Understanding Word Error Rate in Automatic Speech Recognition. I: *Clari.com* [blogg], https://www.clari.com/blog/word-error-rate/. Hämtad 6 februari 2024.

Kloss, Heinz, 1978: Die Entwicklung neuer germanischer Kultursprachen seit 1800. 2:a upplaga. Düsseldorf: Schwann.

Kuronen, Mikko & Leinonen, Kari, 2001: Fonetiska skillnader mellan finlandssvenska och rikssvenska. I: Jönsson, L. m.fl. (red.). *Svenskans beskrivning 24. Förhandlingar vid Tjugofjärde sammankomsten för svenskans beskrivning*. Linköping: Linköping University Electronic Press 006. S. 125–138.

O'Shaughnessy, Douglas, 2008: Automatic speech recognition: History, methods and challenges. *Pattern recognition*, 41(10). S. 2965–2979.

Papastratis, Ilias, 2021: Speech recognition: A review of the different deep learning approaches. I: *The AI Summer*. https://theaisummer.com/speech-recognition/. Hämtad 7 april 2024.

Reuter, Mikael, 1992: Swedish as a pluricentric language. I: Clyne, M. (red.): *Pluricentric Languages. Different Norms in Different Countries*. Berlin/New York: Mouton de Gruyter. S. 101–116.

Reuter, Mikael, 2015a: Finlandssvenskt uttal. I: Tandefelt, M. (red.): *Gruppspråk, samspråk, två språk*. Svenskan i Finland – i dag och i går 1:2. SLS. S. 19–34.

Reuter, Mikael, 2015b: Så här ska det låta. Om finlandssvenska och språkriktighet. Helsingfors: Scriptum.

Sonix, 2024: Automatic speech recognition: A comprehensive guide to ASR technology. I: *Sonix.ai* [blogg], https://sonix.ai/resources/what-asr/. Hämtad 7 april 2024.

Sveriges riksdag. Partiledardebatt 18 juni 2014. https://www.riksdagen.se/sv/webb-tv/video/partiledardebatt/partiledardebatt_h1c120140618pd/. Hämtad 22 februari 2024.

Tillgänglighetskrav.fi. https://www.tillganglighetskrav.fi. Hämtad 25 februari 2024.

Triggerfish.se. En snabbguide till EU:s tillgänglighetsdirektiv. https://www.triggerfish.se/aktuellt/kunskap/en-snabbguide-till-eus-tillganglighetsdirektiv-anpassa-din-verksamhet-nu-och-sitt-sakert-2025/. Hämtad 6 april 2024.

Wang, Ye-Yi, Acero, Alex & Chelba, Ciprian, 2003: Is word error rate a good indicator for spoken language understandign accuracy. I: *2003 IEEE workshop on automatic speech recognition and understanding*. IEEE Cat. No 03EX721. S. 577–582.

Webbriktlinjer. https://www.digg.se/webbriktlinjer. Hämtad 6 april 2024.

Lyhennelmä

Tutkimuksen taustaa

Plurisentriset kielet ovat kieliä, joilla on virallinen asema useammassa kuin yhdessä maassa (Clyne 1992a:1 f.). Monet plurisentriset kielet muodostavat jatkumon, jossa kieltä puhutaan valtiorajojen yli. Tämä on tilanne ruotsin kielen kanssa, ruotsinruotsia puhutaan Ruotsissa ja suomenruotsia Suomessa. Ruotsinruotsi ja suomenruotsi ovat siis ruotsin kielen kaksi kielivarianttia. Tavallista plurisentrisille kielille on, että joku tai jotkin maat nauttivat valta-asemasta. Valta-aseman voi saavuttaa maa, jossa puhutaan vanhinta tai alkuperäisintä kielivarianttia, tai maa, jonka kielivariantin puhujat ovat enemmistössä. Lisäksi valta-aseman voi saavuttaa myös poliittisen tai taloudellisen vahvuuden kautta. Kielivariantit voidaan siis määritellä dominoiviksi (*dominant*) tai ei-dominoiviksi (*non-dominant*). Ruotsinruotsi on ruotsin kielen dominoiva kielivariantti ja suomenruotsi taas ei-dominoiva kielivariantti.

On tavallista, että kielivarianttien välille syntyy ääntämyksellisiä eroja, kun kielivariantteja puhutaan maantieteellisesti eri alueilla. Suomenruotsin ja ruotsinruotsin välillä esiintyy useita ääntämyksellisiä eroja, joista osa on pysyviä, esimerkiksi suomenruotsista puuttuvat aksenttierot ja osa on vapaassa vaihtelussa, puhuja voi esimerkiksi valita, että painottaako hän sanoja suomenruotsin vai ruotsinruotsin mukaisesti (Reuter 2015b:21–31). Muun muassa näiden ääntämyserojen perusteella voimme odottaa, että suomenruotsin ja ruotsinruotsin puheentunnistuksessa ilmenee eroavaisuuksia.

Kieliteknologia on kattotermi, joka käsittää kaikenlaisen tietokoneiden ja kielten vuorovaikutuksen, esimerkiksi kääntämisen ja puheentunnistuksen. Puheentunnistus on yksi kieliteknologian osa-alue ja keskittyy puheen muuntamiseksi tekstimuotoon (Jurafsky ja James 2023). Puheentunnistuksen tuomat teknologiaratkaisut auttavat meitä olemaan tehokkaampia, kun voimme käyttää puhekomentoja esimerkiksi näppäimistöllä kirjoittamisen sijasta. Tehokkuuden tuomien hyötyjen lisäksi on puheentunnistus myös tärkeä osa saavutettavien apukeinojen kehittämisessä. Esimerkiksi henkilöt, joilla on rajoitettuja liikeratoja, voivat saada merkittävää apua puheentunnistuksesta ja äänikomentojen käyttämisestä näppäimistön tai kosketusnäytön sijasta (Tillgänglighetskrav.fi).

Puheentunnistuksessa on kuitenkin joitakin keskeisiä haasteita. Yksi isoimmista haasteista on se, että puheessa on niin paljon vaihtelua. Vaihtelua puheeseen tuo puhujien ikä, sukupuoli,

anatomia, puhenopeus sekä erilaiset murteet (Forsberg 2003:7). Lisäksi taustamelu ja puheen äänittämiseen käytettävät laitteet, kuten mikrofoni, voivat vaikuttaa puheentunnistuksen onnistumiseen (Arora ja Singh 2012:37).

Tutkimuksen tavoite ja tutkimuskysymykset

Tämän tutkimuksen päätavoitteena oli selvittää ilmeneekö suomenruotsin ja ruotsinruotsin puheentunnistuksessa merkittäviä eroja. Tutkin, että mistä puheentunnistuksen tuottamat virhetulkinnat voivat johtua sekä ovatko puheentunnistuksen tuottamat virheet samantyyppisiä molempien kielivarianttien kohdalla ja onko virheiden takana samoja syitä.

Tutkimuksessani vastaan seuraaviin kysymyksiin:

- Miten tämän päivän puheentunnistus käsittelee suomenruotsia verrattuna ruotsinruotsiin?
- Mitkä syyt johtavat puheentunnistuksen virhetulkintoihin suomenruotsissa ja ruotsinruotsissa?
- Mitä mahdollisuuksia puheentunnistuksen kehittämiselle on, jotta voisimme paremmin kohdata suomenruotsia puhuvien tarpeet?

Ensimmäistä kysymystä varten puheentunnistustyökalun tekemiä virheitä verrattiin kielivarianttien, suomenruotsin ja ruotsinruotsin, välillä. Toinen kysymys käsittelee niitä tekijöitä, jotka ovat voineet johtaa puheentunnistustyökalun virhetulkintoihin, esimerkiksi ääntämysmallit tai leksikaaliset erot. Puheentunnistuksen tekemien virheiden tunnistaminen ja niihin johtavien syiden analysointi johdattaa meidät kolmanteen kysymykseen ja antaa pohjan keskustelulle, miten puheentunnistustyökaluja voitaisiin edelleen kehittää.

Aineisto ja tutkimusmenetelmät

Tutkimukseni aineisto koostuu valmiista äänitiedostoista. Suomenruotsia koskevat äänitiedostot ovat kerätty Alto-yliopiston puheentunnistuskorpuksesta ja sisältävät eduskunnan istuntojen puheenvuoroja. Kokosin kaikki syyskauden puheenvuorot vuodelta 2014, jonka jälkeen leikkasin niistä pois kaikki suomenkieliset puheenvuorot ja kokosin jäljellejäävät puheenvuorot uusiksi äänitiedostoiksi. Nämä äänitiedostot syötin puheentunnistustyökaluun. Suomenruotsin aineisto koostui kaikenkaikkiaan 10 932 sanasta ja

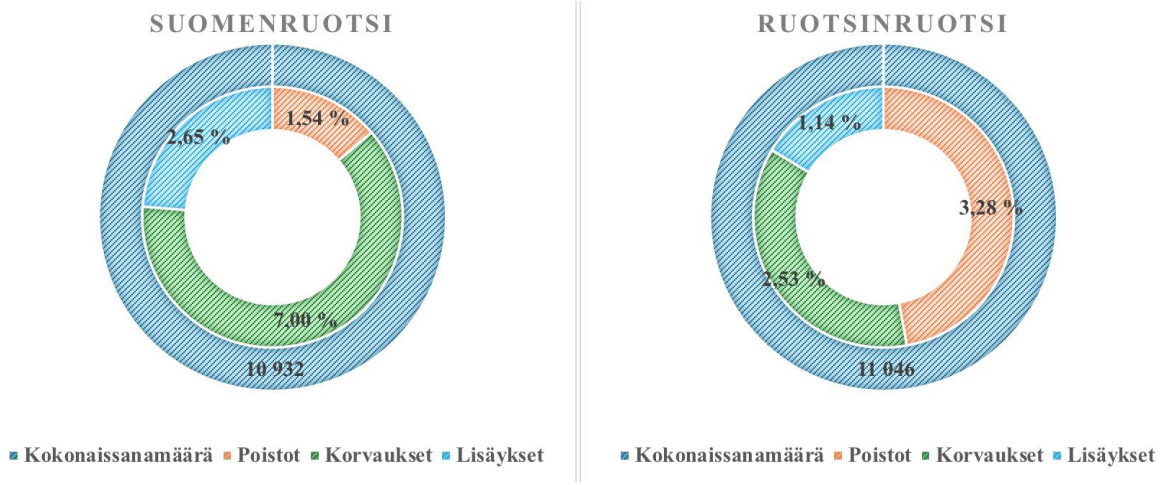
se vastasi 88:aa puhuttua minuuttia. Ruotsinruotsin äänistiedostot kokosin Ruotsin eduskunnan verkkosivuilta ja materiaaliksi valikoitui puoluejohtajien keskustelu samalta vuodelta 2014. Ruotsinruotsia koskevasta materiaalista ei tarvinnut leikata mitään pois vaan se oli sellaisenaan valmista puheentunnistustyökaluun syötettäväksi, sillä saatoin olettaa, että kaikki puhujat puhuvat ruotsia äidinkielenään. Ruotsinruotsin aineisto koostui 11 046 sanasta, mikä vastasi 68:aa minuuttia puhetta. Molempien kielivarianttien kohdalla materiaalin ohessa oli saatavilla transkriptio, mitä tarvittiin puheentunnistuksen onnistumisen arvioimisessa.

Nämä äänitiedostot syötettiin puheentunnistustyökaluun, jonka jälkeen puheentunnistusta verrattiin systemaattisesti alkuperäisiin transkriptioihin. Tällä tavoin puheentunnistuksen tekemät virheet tunnistettiin sekä jaoteltiin kolmeen WER-virhekategoriaan (*word error rate*), joita ovat korvaukset, lisäykset ja poistot. Korvaukset ovat virheitä, joissa puheessa esiintyvä sana on korvattu jollakin toisella sanalla. Poistot ovat virheitä, joissa puheentunnistus on jättänyt joitakin sanoja täysin tunnistamatta eli puheentunnistuksen tulkinnasta puuttuu sanoja. Lisäykset taas ovat virheitä, jotka puheentunnistus on lisännyt eli ylimääräisiä sanoja puheentunnistuksen tulkinnassa.

Virheiden tunnistamisen pohjalta jokaiselle äänitiedostolle laskettiin uniikki WER-virheprosentti, jossa virheiden määrä jaetaan kyseessä olevan äänitiedoston kokonaissanamäärällä. Tämän jälkeen WER-virhekategorioita pystyttiin analysoimaan. Analyysissa keskityttiin tutkimaan ovatko virhekategoriat yhtä suuria molemmissa kielivarianteissa, millaisia virheitä kuhunkin kategoriaan sisältyy ja ovatko virheet samoja kielivarianttien kesken. Lopuksi tunnistettiin mahdollisia syitä, jotka ovat johtaneet näihin virheisiin kussakin kategoriassa ja edelleen, koskivatko samat syyt molempia kielivariantteja.

Tulokset ja pohdinta

Yleiskatsaus virhekategorioista kuvataan kaaviossa 1. Ulompi rinkula kaaviossa osoittaa kokonaissanamäärän kummassakin kielivariantissa ja sisemmät rinkulat osoittavat kunkin virhekategorian prosentuaalisen osuuden kokonaissanamäärään suhteutettuna.



Kaavio 1. Virhekategoriat prosenttiosuuksina kokonaissanamäärään suhteutettuna.

Kategoria korvaukset on suomenruotsin isoin kategoria tasan seitsemällä prosentilla ja ruotsinruotsin toisiksi isoin kategoria hiukan alle kolmella prosentilla. Kategoria pitää molemissa kielivarianteissa sisällään hyvin monenlaisia virheitä, osa sanoista on korvattu täysin eri sanalla ja osassa virhetulkinnoista on kyse sanan korvaamisesta väärällä taivutusmuodolla. Kategoria pitää siis sisällään hyvin paljon virhetulkintoja, mutta keskimäärin suuri osa virheistä esiintyvät vain kerran. Suomenruotsin materiaalissa esiintyviin korvauksiin johtavia syitä olivat puhujien ikä ja artikulaatio, jolloin puheentunnistus ei onnistunut tunnistamaan puhuttuja sanoja. Samat piirteet nousivat esille myös ruotsinruotsin materiaalissa, mutta positiviisessa valossa. Ruotsalaiset puhujat olivat siis keskimäärin nuorempia ja puhuivat selkeämmin, mikä johti parempiin tuloksiin. Yksi yhteinen virheisiin johtava syy oli sanojen sulautuminen yhteen, esimerkiksi frsaain *har det* tulkinta sanaksi *hade*. Suomenruotsin materiaalista nousi esille myös yksi erityinen piirre, jossa osa virhetulkinnoista näytti johtuvan ääntämisestä. Tästä esimerkkinä on tapaus, missä suomenruotsin materiaalissa esiintyvä sana *talman* oli tulkittu mm. sanaksi *dahlman*. Tämän voi selittää sillä, että suomenruotsin soinniton klusiili [t] lausutaan tavallisesti ilman voimakasta uloshengitystä, toisin kuten ruotsinruotsissa on tapana, jolloin se saattaa kuulostaa konsonantin soinnilliselta vastaparilta [d].

Kategoria poistot on suomenruotsin pienin kategoria hieman alle kahdella prosentilla ja ruotsinruotsin isoin kategoria yli kolmella prosentilla. Vaikka kategoria on hyvin erikokoinen kielivarianttien kesken, yhteistä on se, että kategoria ei keskity mihinkään tiettyyn sanaluokkaan. Molempien kielivarianttien poistoja yhdisti se, että pikkusanat, kuten

prepositiot, useasti sulautuivat niitä ympäröiviin sanoihin. Kategoria oli huomattavasti isompi ruotsinruotsin materiaalissa ja siellä korostui pikkusanojen lisäksi myös yllättävän pitkien sanojen poistot, kuten *egentligen*, *någonting* ja *Socialdemokraterna*. Tähän isoin vaikuttava tekijä oli ruotsalaisten huomattavasti nopeampi puhetahti verrattuna suomenruotsalaisiin.

Kategoria lisäykset on suomenruotsin toiseksi isoin kategoria hieman alle kolmella prosentilla ja ruotsinruotsin pienin kategoria vain hieman yli yhdellä prosentilla. Tässä kategoriassa yhteistä kielivarianteille on se, että kategoria koostui lähinnä pikkusanoista, kuten prepositioista ja pronomeneista. Useiten lisätty sana, *och*, oli sama molemmissa kielivarianteissa ja se johtuu todennäköisesti siitä, että sana lisättiin niihin kohtiin, joissa oikeastaan olisi kuulunut olla piste eli merkitsemään lauseiden loppuja. Kategorian suuruus suomenruotsin materiaalissa johtuu siitä, että nämä virheet olivat usein yhteydessä korvauksiin, joka oli suomenruotsin isoin kategoria. Muut ruotsinruotsin materiaalissa esiintyvät virheet olivat vaikeammin tulkittavia, mutta todennäköisiä syitä virhetulkintoihin olivat puhujien kova puhenopeus ja sanojen lyhentäminen sekä assimilaatio.

Suomenruotsin keskimääräinen WER-virheprosentti oli 0,14 % ja ruotsinruotsin vastaava luku oli 0,07 %. Voidaan siis todeta, että ruotsinruotsin puheentunnistuksen tulokset olivat kaksi kertaa tarkempia ja parempia, kuin suomenruotsin. On siis selkeä ero siinä, miten työkalu tunnistaa suomenruotsia ja ruotsinruotsia. Tulokset vihjailevat siitä, että eroavaisuudet eivät johdu plurisentrisille kielille ominaisista piirteistä, kuten ääntämyseroista, vaan todennäköisempi syy on puheentunnistustyökalun opettamiseen käytetyn harjoitusmateriaalin määrä. On todennäköistä, että harjoitusmateriaalia, eli puhetta, on helpompi kerätä kielelle tai kielivariantille, jolla on paljon puhujia. Näin ollen olisi järkeenkäypää, että ruotsinruotsia kohden on käytetty enemmän harjoitusmateriaalia.

Erilaisten apuvälineiden ja teknologiaratkaisujen yleistyessä on tähän kiinnittettävä huomiota. Meidän tulisi löytää ratkaisuja siihen, miten voimme taata luotettavan datankeruun myös kielille tai kielivarianteille, joilla on vähän puhujia. Tällä tavoin voisimme taata, että myös kielten ei-dominoivien kielivarianttien puhujat voivat tulevaisuudessa hyödyntää toimivia puheentunnistustyökaluja.