Från dialektinspelning till talspråkskorpus – beskrivning av ett korpusbygge

Lisa Södergård, Svenska litteratursällskapet i Finland
& Therese Leinonen, Åbo universitet

Abstract
The Talko corpus of Swedish spoken in Finland is a new research tool consisting of audio files linked to annotation, i.e., transcriptions on two parallel levels and part-of-speech tagging. The corpus is searchable through a web-based interface. The recordings were made in 2005–2008 in all parts of Swedish-language Finland. They have been transcribed in a broad phonetic transcription as well as in a standard orthographic transcription. The part-of-speech tagging is done with TreeTagger, trained on the Stockholm-Umeå Corpus of written Swedish. The automatically produced part-of-speech tags are manually corrected for subsets of the data, and the manually corrected data are subsequently added to the training data. This will gradually improve the result of the automatic tagging and compensate for differences between spoken and written Swedish and between Finland-Swedish and Sweden-Swedish.

1 Inledning

I augusti 2014 öppnades den finlandssvenska talspråkskorpusen Talko[1] för användarna. Det primära materialet i korpusen utgörs av ljudfiler men en förutsättning för att materialet ska bli sökbart är att det finns olika former av annotering. Ljudfilerna i Talko har försetts med annotering i form av två olika typer av utskrifter eller transkriptioner. Dessutom har orden i transkriptionen försetts med uppgifter om ordklass. I korpusgränssnittet Glossa sammanförs all information och blir sökbar för korpusanvändaren.

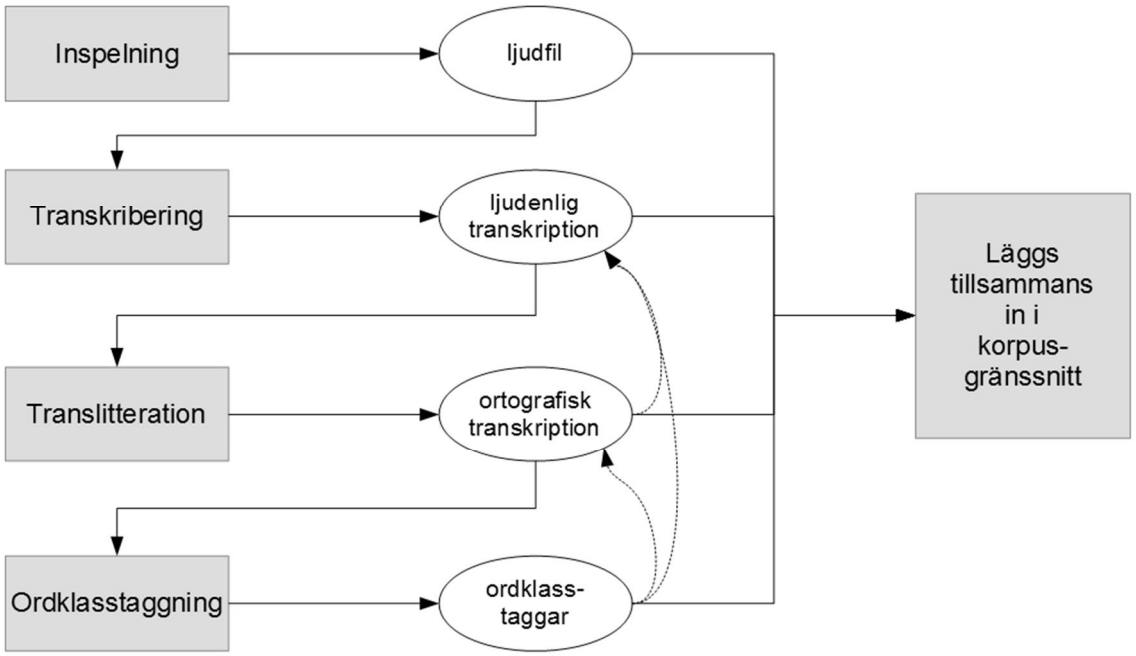
Talko är tillgänglig via webben och får användas för vetenskapliga ändamål. Korpusen upprätthålls av Svenska litteratursällskapet i Finland i samarbete med Tekstlaboratoriet vid Universitetet i Oslo. En länk till korpusen och utförligare information om innehåll och användningsmöjligheter finns på www.sls.fi/talko.

I den här studien beskriver vi olika arbetsmoment som har krävts för att utifrån talspråksinspelningar bygga upp en sökbar korpus. En översikt över arbetsmomenten ges i figur 1. För många av arbetsmomenten har färdiga lösningar från andra motsvarande projekt kunnat användas eller anpassas. I vissa fall uppstår problemställningar som är mer specifika för det finlandssvenska materialet.

2 Material

I korpusen ingår inledningsvis inspelningar gjorda inom projektet *Spara det finlandssvenska talet* (kort Spara talet), som genomfördes vid Svenska litteratursällskapet under åren 2005–2008. I Talko 0.1, som öppnades i augusti 2014, ingick avsnitt ur 100 intervjuer med 115 äldre och yngre personer från olika delar av Svenskfinland. I merparten av inspelningarna medverkar en intervjuare och en informant, i en del inspelningar intervjuas två informanter samtidigt. Inspelningarna

[1] Ordet *talko* används i finlandssvenskan om arbete som utförs frivilligt tillsammans.



Figur 1. Arbetsflöde. Korpusmaterialet består av fyra olika parallella lager, vilka kräver var sitt arbetsmoment. Arbetet med korpusen innebär återkoppling mellan de olika lagren eftersom dessa påverkar varandra så att t.ex. antalet ord måste vara detsamma i varje lager.

har valts ut med avsikt att få spridning för variablerna hemort, kön och ålder. Också inspelningens ljudkvalitet och innehåll[2] har haft en viss inverkan vid valet av material för korpusen. Intervjuerna i sin helhet är 40–60 minuter långa men för korpusen har ca 20 minuter ur varje inspelning valts ut[3] och transkriberats. För en noggrannare beskrivning av insamlingsmetoder och material för Spara talet se Ivars & Södergård (2007, 2009).

3 Transkriptioner

Det material som ingår i Talko har transkriberats på två olika nivåer parallellt: både med en ljudenlig transkription och med en ortografisk transkription. Tack vare detta förbättras sökmöjligheterna avsevärt. Man kan t.ex. göra sökningar i den ortografiska transkriptionen utan att behöva känna till alla olika uttal av ett ord. Samtidigt ligger den ljudenliga transkriptionen närmare det som verkligen sägs på inspelningarna.

Att materialet i Talko annoteras automatiskt med hjälp av en taggare ställer också krav på transkriptionen. Ett av de mest synliga är att varje ord måste stå separerat av mellanslag i utskriften eftersom taggaren tilldelar varje token information om ordklass. Ett token är den minsta enhet som används vid annoteringen av en korpus. Token omges i regel av mellanslag eller av mellanslag och skiljetecken. Grafiska ord utgör token, men även skiljetecken kan räknas som token och i Talko utgör

[2] Avsnitt innehållande känsliga personuppgifter har inte tagits med.
[3] Vanligen ett avsnitt börjande ca 10 minuter in i inspelningen och fram till 30 minuter.

t.ex. även pausteckнen token. Ett exempel på hur detta påverkar transkriptionen är frasen *det är* som ofta uttalas som ett enda långt *de:* (jfr Forsskåhl 2009). För att taggningen ska bli *pronomen + verb* måste transkriptionerna vara *de e* respektive *det är* också för uttalet *de:*. En del sammanskrivningar med bindestreck förekommer ändå i utskriften och tolkas då som ett enda token. Det handlar om token som kan taggas som en enhet, t.ex. *det-här, de-där* eller *sådan-här* som tolkas som pronomen. I det senare fallet förekommer uttalsformerna *sånhäna* och *sånhä:nt* som motiverar sammanskrivningarna *sådana-här* och *sådant-här*.

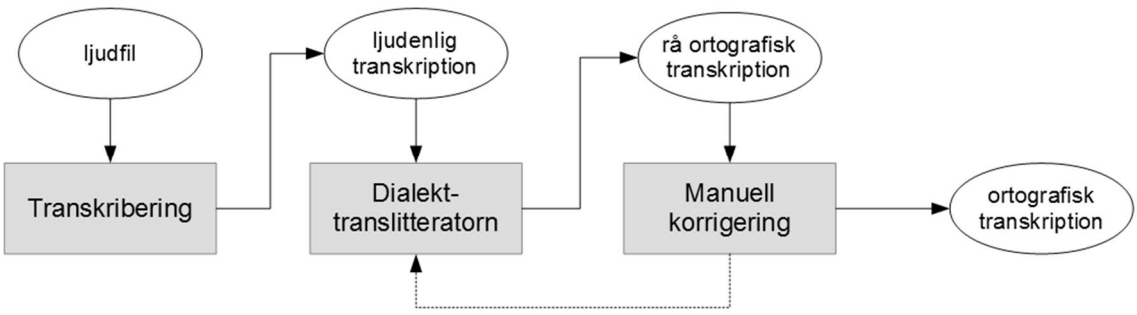
Ljudenlig transkription

Den ljudenliga transkriptionen som ingår i Talko är ett försök att gestalta hur det som sägs på inspelningen låter. Den baserar sig inte på en systematisk inventering av fonemsystemen i de olika dialekterna (jfr Wiik & Östman 1983: 184) utan kan närmast beskrivas som ett slags grov fonetisk utskrift med alfabetets vanliga bokstäver och ett antal bokstavskombinationer, t.ex. *sj* för sje-ljudet och *tj* för tje-ljudet. Kolon används för att markera långa vokaler, jfr transkriptionen *bara* för uttal med kort *a* och *ba:ra* för uttal med långt *a* av ordet *bara*. Långa konsonanter dubbelteckнas, t.ex. *tykkas* för ordet *tyckas*. Alla fonem noteras inte, t.ex. kakuminalt l-ljud har inte markerats i den ljudenliga transkriptionen.

Att göra transkriptioner av inspelningar är ett krävande arbete och det är också känt att transkriberarens egen varietet påverkar tolkningen av inspelningen (Benson 1982: 156). Det är många personer som gjort de transkriptioner som ingår i Talko. Huvudregeln har varit att inspelningarna transkriberats av en person med stor förtrogenhet med de varieteter som talas i området.

Ortografisk transkription

Utgående från den ljudenliga transkriptionen skapas en ortografisk transkription (jfr Benson 1982: 155). Som hjälp vid överföringen används en dialekttranslitterator som utvecklats vid Tekstlaboratoriet i Oslo (Johannessen et al. 2009: 75). Translitteratorn gör en första råversion som sedan korrigeras manuellt, se översikt över arbetsflödet i figur 2. Överföringen till den ortografiska transkriptionen sker ord för ord,



Figur 2. Överföringen från ljudenlig till ortografisk transkription. Manuella rättelser som görs i den ortografiska transkriptionen återkopplas till dialekttranslitteratorn, vilket förbättrar den automatiska translitterationen av följande fil.

och resultatet motsvarar alltså fortsättningsvis det som sägs på inspelningen men varje ord anges i skriftspråklig form.

Svenska Akademiens ordlista (SAOL) utgör huvudsaklig källa för den ortografiska utskriften. I transkriptionerna förekommer naturligtvis också ord som inte ingår i SAOL. Det handlar huvudsakligen om tre olika typer av ord: dialektala ord, finlandismer och utländska ord. För dialektord och finlandismer används *Ordbok över Finlands svenska folkmål* (FO) och *Finlandssvensk ordbok* (af Hällström-Reijonen & Reuter 2008) som utgångspunkt för den ortografiska formen. FO i sin tur använder *Svenska akademiens ordbok* som utgångspunkt när det är möjligt och för övriga ord konstrueras en egen uppslagsform (Slotte 1981: 51f). Uppslagsformen i FO kan i många fall användas för att konstruera den ortografiska transkriptionen i Talko: i FO finns uppslagsformen *mära*, som är ett verb av första konjugationen, och presensformen *mä:rar* i den ljudenliga transkriptionen i Talko får därmed den ortografiska transkriptionen *märar*. Ord som inte ingår i SAOL förses med taggen ”x” och blir därmed också sökbara.

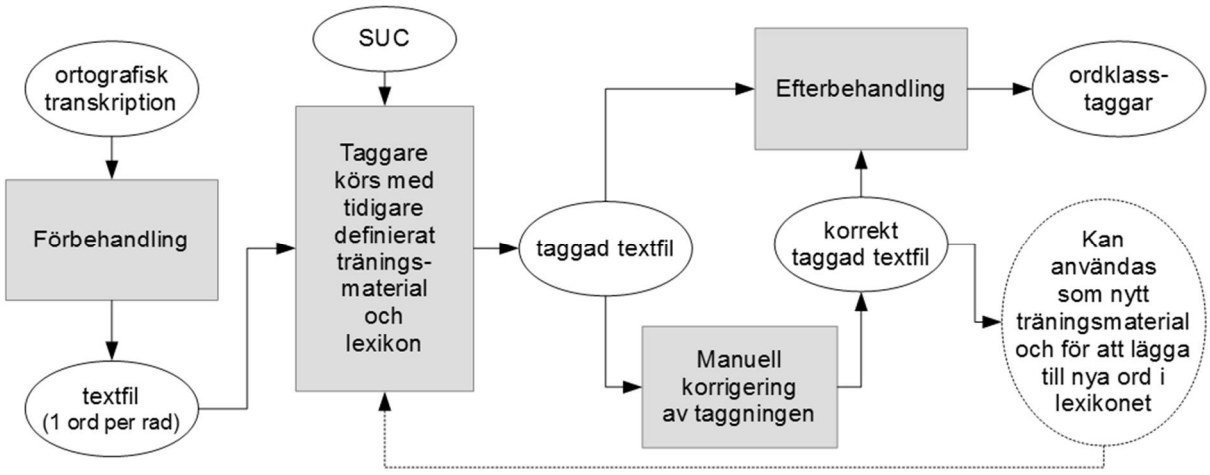
Inte bara ortografin utan även bl.a. verbböjningen normaliseras. Det innebär att den rika variation i både uttal och böjningsformer som finns i den ljudenliga utskriften förenhetligas enligt standardsvenskt mönster:

	Ljudenlig transkription	Ortografisk transkription
Infinitiv	*le:sa, le:ssa, lesa, les*	läsa
Presens	*le:sär, lesär, le:ssä:*	läser
Preteritum	*lest, le:st, le:ste, las*	läste
Supinum	*le:st, lesi, le:si*	läst

I enlighet med SAOL blir de i finlandssvenskan vanliga uttrycken *sku bo:rda* och *sku måsta* i den ortografiska utskriften *skulle böra* respektive *skulle måsta*. SAOL anger infinitivformen *måsta* med beteckningarna provinsiellt och finlandssvenska.

I all transkription ingår ett visst mått av tolkning. I den ortografiska transkriptionen kan t.ex. sammanfall mellan uttal av olika böjningsformer vara svåra att tolka. Detta gäller t.ex. verb av första konjugationen där alla böjningsformer (infinitiv, presens, preteritum och supinum) kan låta lika, exempelvis verbet *tala* med uttalet *ta:la* i alla böjningsformer. Kontexten i inspelningen blir då avgörande för vilken tempusform som väljs i den ortografiska transkriptionen.

I några fall har ett flertal olika uppslagsformer som ingår i FO sammanförts till en enda i den ortografiska transkriptionen i Talko. Detta gäller t.ex. adverbet *här* som i Talko används för bland annat uttalen *hä:r, he:r, jä:r, sjär, sjier, sjenn, hije:, ije:, hijenn, ije:nan.* I FO upptas de under uppslagsorden *här, härna, hithjär, hithjärnan*. På samma sätt har pronomenet *he* (i FO uppslagsordet *het)* i Talko sammanförts med *det*. Trots denna förenkling i den ortografiska utskriften är det möjligt att hitta alla *he* tack vare möjligheten att kombinera sökningar på uttal och ortografi.



Figur 3. Arbetsflöde över ordklasstaggningen. Transkriptionerna förbehandlas för att få ett textformat som taggaren kan läsa. Träningsmaterialet ökar successivt allt eftersom Talkomaterialet korrigeras och det automatiska taggningsresultatet därmed förbättras. Vid efterbehandlingen ges filerna ett format som kan läsas av korpusgränssnittet.

4 Ordklasstaggning

Ordklasstaggningen av Talko görs med den statistiska taggaren TreeTagger (Schmid 1994, 1995) och till att börja med med den manuellt taggade korpusen SUC 3.0 (*Stockholm-Umeå Corpus*, Gustafson-Capková & Hartmann 2006) som träningsmaterial. Träningsmaterialet behövs för att taggaren ska lära sig sannolikheter för vilken typ av ord som brukar höra till en viss ordklass och hur olika taggar kan kombineras syntaktiskt i det språk som ska taggas.

SUC består av svenskt skriftspråk från 1990-talet och används som standard vid träning och utvärdering av ordklasstaggare på svenska. Taggsuppsättningen består av ordklasstaggar med viss morfologisk information (Ejerhed et al. 1992). När en taggare som är tränad på skriftspråk från Sverige körs på finlandssvenskt talspråk kommer taggaren att göra vissa systematiska feltaggningar. En stor del av språket är ändå gemensamt och därför fungerar denna metod väl som ett första steg i taggningsarbetet.

Det automatiska taggningsresultatet kan utvärderas genom att man först taggar ett delmaterial automatiskt och därefter korrigera feltaggningarna vid en manuell genomgång av resultatet. Det handkorrigerade materialet kan dessutom läggas till i träningsmaterialet, vilket avsevärt förbättrar taggningsresultatet för följande delmaterial som ska taggas; se figur 3. Ett motsvarande tillvägagångssätt användes bl.a. vid taggning av talspråk från Oslo med en taggare som inledningsvis tränats på norskt skriftspråk (Søfteland & Nøklestad 2008).

Förbättring av taggningsresultatet genom manuell korrigering

När Talkomaterialet taggas med SUC som träningsmaterial ligger korrektheten under 80 procent. När ca 20 000 token Talkomaterial ingår i träningsmaterialet stiger korrektheten till kring 93 procent (Leinonen 2015). En stor förbättring inträder när taggaren lär sig använda taggarna för diskurspartiklar, tvekljud och uppbackningar

samt pauser som saknas i skriftspråket och som införts i Talko. Övriga taggningar där en stor förbättring sker är infinitivmärke, egennamn, subjunktioner och pronomen. Att en stor förbättring sker vid infinitivmärket beror på att *till* används som infinitivmärke i finlandssvenska dialekter. Dessa *till* blir konsekvent feltaggade när taggaren endast tränats på SUC. Subjunktionernas korrekthet stiger när taggaren av Talkomaterialet bl.a. lär sig att *före* i finlandssvenska ofta är en subjunktion.

När egennamn blir feltaggade blir de i regel taggade som substantiv. Att egennamnen blir feltaggade beror framför allt på att de finlandssvenska ortnamnen som ofta återkommer i intervjuerna saknas i SUC. Ett enkelt sätt att öka korrektheten vore därför att helt lägga in en lista över finlandssvenska ortnamn i det lexikon taggaren använder och som till att börja med byggts upp ur bara SUC.

Att pronomenen i början till stor del blir feltaggade beror på de talrika sammansättningarna med bindestreck (*det-här, de-där, sådan-här* etc.; se ovan) som saknas i SUC. Ett enskilt ord där en stor förbättring sker när Talkomaterial läggs till i träningsmaterialet är *nå,* som är en frekvent turinledande diskurspartikel i finlandssvenska. Innan taggaren tränats på finlandssvenskt talspråk taggas de flesta förekomster felaktigt som verb.

Genus

Substantiv, adjektiv, pronomen och particip markeras i SUC förutom för ordklass också för numerus, genus, kasus och species. Taggsuppsättningen utgår från ett tvågenussystem med utrum och neutrum. I Finland finns dock svenska dialekter med bevarat tregenussystem och dialekter som saknar genus. Vi har valt att ändå följa SUCs uppdelning i utrum och neutrum och tagga maskulina och feminina ord som utrum. Vissa problem uppstår trots det vid markeringen av genus, vilket bl.a. har att göra med skillnader mellan svenskan i Finland och svenskan i Sverige. Här följer ett konkret exempel.

SUC och SAOL ger endast genus neutrum för ordet *lekis*. Informanten helsingfors_ow03 i Talko använder först detta ord i obestämd form singular och det finns inga kongruerande ord i kontexten som skulle avslöja vilket genus ordet har. Taggaren taggar ordet som neutrum i enlighet med SUC. Senare under intervjun använder samma informant ordet i bestämd form *lekisen*, vilket visar att ordet har genus utrum i informantens språkbruk. Ordet *lekis* saknas i *Finlandssvensk ordbok* men det till både form och betydelse nära besläktade *dagis* finns upptaget i ordboken med en kommentar om att ordet är neutrum, vilket implicerar att det inte alltid används som ett neutrum i finlandssvenskan. Ändelsen *-en* i *lekisen* gör att ordet taggas korrekt som ett utrum trots att formen saknas i taggarens träningsmaterial. Vid denna typ av ord uppstår frågan hur man ska göra med det första belägget där genus inte framkommer. Och vad man hade gjort om det andra belägget inte hade funnits hos informanten.

Lånord och kodväxling

I tagguppsättningen ingår en tagg för utländska ord (UO) som i SUC främst används i citat och verktitlar (Ejerhed et al. 1992: 13). I finlandssvenskt talspråk ingår en stor del mer eller mindre integrerade lånord från både finska och engelska. Lånorden rör

sig på en skala mellan morfologiskt helt integrerade lån som utgör etablerade finlandismer (t.ex. *halaren* ’overallen’, *kaverin* ’kompisar’, *keikkor* ’grejor, gig’, *kivogare* ’roligare’) och rena anföringar som i följande exempel:

> *olika kemikalier öh det som på finska kallades (.)* ***tulikukka***(=handelsnamn för rent svavel, svavelblomma) (åbo_om04)

Vid de integrerade lånorden kan man utan problem ge orden fullständiga morfologiska taggar. I fallet *tulikukka* är det naturligt att använda taggen UO. Men gränsdragningen mellan etablerade och tillfälliga lån har visat sig vara svår i materialet och alltid går det helt enkelt inte att veta hur etablerat ett visst lånord är i en viss språkgemenskap. Vid taggningen leder detta till svårigheter i synnerhet med de morfologiska taggarna. I följande två exempel har lånorden inte anpassats morfologiskt:

> *dylika ganska enkla (.)* ***messages***(=meddelanden) (åbo_om04)

> *vi är båda också sådana att eh (.) på det sättet (.)* ***erakoita***(=eremiter) (åbo_ym19)

Bägge orden kan taggas som substantiv, vilket är mer informativt för korpusen än taggen UO. Bägge kan också taggas som plural trots att de behållit pluraländelsen från sina respektive ursprungsspråk. Men att tagga orden för genus är betydligt svårare. Ett val måste i dessa fall göras mellan antingen taggning med den grammatiskt icke-informativa taggen UO eller med delvis ofullständiga SUC-taggar, och gränsdragningen mellan dessa alternativ har visat sig vara svår.

5 Slutsatser

Genom transkriptioner och taggningar görs ljudinspelningarna i Talko sökbara på ett mångsidigt och varierat sätt. Det är t.ex. möjligt att göra sökningar på ortografiska former, på uttal (dvs. i ljudenliga utskriften), på ordklasser eller rentav på en kombination av dessa. Trots att materialet i Talko i alla arbetsskeden genomgår flera bearbetningar och kontroller förekommer fel och svårtolkade fall. Det är därför av yttersta vikt att korpusanvändaren själv kan gå in och lyssna på det primära materialet, inspelningen.

Den första publicerade korpusversionen Talko 0.1 är att betrakta som ett slags demoversion. Utifrån bl.a. användarrespons kommer materialet att bearbetas ytterligare och vissa ändringar göras till Talko 1.0 som lanseras inom 2015. I Talko 0.1 ingick ungefär en tredjedel av det planerade Spara talet-materialet som kommer att läggas in i korpusen efter hand. I Talko 1.0 ingår mer Spara talet-material och dessutom 29 avsnitt ur äldre arkivinspelningar, som även ingår i Harling-Kranck (1998). För att förbättra resultatet av den automatiska taggningen har mer material korrigerats manuellt så att andelen finlandssvenskt talspråk i taggarens träningsmaterial ökar. Dessutom har hela den förteckning över ortnamn som ingår i *Svenska ortnamn i Finland* (Mattfolk & Vidberg 2012) lagts till träningsmaterialet, vilket innebär att taggningen av ortnamn förbättras avsevärt.

Referenser

Benson, Sven, 1982. Utskrift på olika nivåer. I *Talspråksforskning i Norden. Mål–material–metoder* (152–163), red. av Mats Thelander. Lund: Studentlitteratur.

Ejerhed, Eva, Gunnel Källgren, Ola Wennstedt & Magnus Åström, 1992. *The linguistic annotation system of the Stockholm-Umeå corpus project*. [Report 33] Umeå universitet: Department of General Linguistics.

FO = *Ordbok över Finland svenska folkmål*, 1976–. Helsingfors: Forskningscentralen för de inhemska språken/Svenska litteratursällskapet i Finland.

Forsskåhl, Mona, 2009. *Konstruktioner i interaktion.* de e *som resurs i samtal*. [Studier i nordisk filologi 83] Helsingfors: Svenska litteratursällskapet i Finland.

Gustafson-Capková, Sofia & Britt Hartmann, 2006. Manual of the Stockholm Umeå Corpus version 2.0 . <http://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf>

Harling-Kranck, Gunilla, 1998. *Från Pyttis till Nedervetil. Tjugonio dialektprov från Nyland, Åboland, Åland och Österbotten*. Helsingfors: Svenska litteratursällskapet i Finland.

af Hällström-Reijonen, Charlotta & Mikael Reuter, 2008. *Finlandssvensk ordbok*. Helsingfors: Schildts.

Ivars, Ann-Marie & Lisa Södergård, 2007. Spara det finlandssvenska talet. I *Nordisk dialektologi og sociolingvistik* (202–206), red. av Torben Arboe. Århus: Peter Skautrup Centret för Jysk dialektforskning.

Ivars, Ann-Marie & Lisa Södergård, 2009. Spara det finlandssvenska talet, slutrapport. URL http://urn.fi/URN:NBN:fi-fe201004191691.

Johannessen, Janne Bondi et al., 2009. The Nordic Dialect Corpus – An advanced research tool. I *NODALIDA 2009 Conference Proceedings* (73–80), red. av Kristiina Jokinen & Eckhard Bick. [NEALT Proceedings Series 4].

Leinonen, Therese, 2015. Ordklasstaggning av finlandssvenskt talspråk. I *Talet lever! Fyra studier i svenskt talspråk i Finland*. [Studier i nordisk filologi 86] Helsingfors: Svenska litteratursällskapet i Finland.

Mattfolk, Leila & Maria Vidberg (red.), 2012. *Svenska ortnamn i Finland*. [Institutet för de inhemska språkens webbpublikationer 32] Helsingfors: Institutet för de inhemska språken.

SAOL = *Svenska Akademiens ordlista över svenska språket*, [13:e uppl.] 2006. Stockholm: Svenska Akademien.

Schmid, Helmut, 1994. Probabilistic part-of-speech tagging using decision trees. I *Proceedings of the International Conference on New Methods in Language Processing* (255–261). Manchester: UMIST.

Schmid, Helmut, 1995. Improvements in part-of-speech tagging with an application to German. I *Proceedings of the ACL SIGDAT-Workshop* (47–50). Dublin: ACL.

Slotte, Peter, 1981. Ordbok över Finlands svenska folkmål – problem och metoder. I *De finlandssvenska dialekterna i forskning och funktion* (45–64), red. av Bengt Loman. [Meddelanden från stiftelsens för Åbo Akademi forskningsinstitut nr 64] Åbo: Åbo Akademi.

Søfteland, Åshild & Anders Nøklestad, 2008. Manuell morfologisk tagging av NoTa-materialet med støtte fra en statistisk tagger. I *Språk i Oslo: Ny forskning omkring talespråk* (226–234), red. av Janne Bondi Johannessen & Kristin Hagen. Oslo: Novus.

Wiik, Barbro & Jan-Ola Östman, 1983. Skriftspråk och identitet. I *Struktur och variation. Festskrift till Bengt Loman* (181–216), red. av Erik Andersson, Mirja Saari & Peter Slotte. [Meddelanden för Stiftelsens för Åbo Akademi forskningsinstitut nr 85] Åbo: Åbo Akademi.