SLS SVENSKA LITTERATURSÄLLSKAPET I FINLAND

Talko – korpus över den talade svenskan i Finland

Information till användare

Innehåll

Inledning

Talspråkskorpusen Talko är Svenska litteratursällskapets verktyg för att göra arkivinspelningar tillgängliga för forskning och undervisning. Korpusen innehåller ljudinspelningar och tillhörande sökbara utskrifter. Utskrifterna har försetts med ordklasstaggar och en del morfologisk information, vilket utökar sökmöjligheterna ytterligare. Talko innehåller huvudsakligen inspelningar gjorda 2005–2008 inom projektet Spara det finlandssvenska talet (Spara talet) och endast en liten andel äldre arkivinspelningar.

Korpusen blev en naturlig fortsättning på den omfattande insamling som SLS arkiv gjorde i samband med Spara talet. Planeringen av korpusen inleddes redan 2008 och de första utskrifterna gjordes efter det. Den första versionen av korpusen, Talko 0.1 publicerades hösten 2014 och sedan dess har korpusen uppdaterats vid ett flertal tillfällen. När Talko version 3.0 lanserades våren 2021 nåddes slutligen det uppställda målet, d.v.s. att ca 10 % eller drygt 100 timmar av Spara talets totalt ca 1000 h inspelat ljudmaterial skulle ingå i korpusen.

Arbetet med korpusen har pågått i drygt 10 år och involverat sammanlagt närmare 30 personer, i kortare perioder. På arkivet har 1 - 2 arkivarier arbetat med korpusen, som en uppgift bland andra arbetsuppgifter. För utvecklandet av den automatiska taggningen och för hjälp med automatiserandet av arbetsprocesserna anlitades utomstående expertis. För transkriberingen anlitades drygt 20 olika personer. Avsikten med att engagera så många olika transkriberare var att varje person skulle skriva ut inspelningar där dialekterna och varieteterna låg nära det egna språket.

Vartefter arbetet med utskrifterna och den automatiska taggningen framskred, modifierades riktlinjerna för transkriptionerna. En effekt av att arbete med korpusen pågått lång tid och involverat många personer är att det även förekommer en del inkonsekvenser. Att utskrifterna kunde annoteras automatiskt med bl.a. ordklasser och morfologisk information möjliggjorde att stora mängder material kunde bearbetas, men innebär samtidigt att det förekommer felaktigheter.

För korpusanvändarens del är det viktigt att göra sökningar på olika sätt för att verifiera att sökningen verkligen gett de träffar man är intresserad av. Centralt är också att användaren själv lyssnar på inspelningen. Talspråkskorpusen Talko ger möjlighet att utforska svenskan som talas i Finland ur många olika synvinklar.

Registrering och inloggning

Inloggningen sker via webbssidan www.sls.fi/talko.

Om du har en e-postadress från ett universitet kan du logga in i Talko direkt via CLARIN-inloggningen. Övriga användare kan ansöka om ett eget CLARIN-användarkonto.

Sökningar

Sökningar i Talko görs i användargränssnittet Glossa. Söktips finns i användarhandledningen.

Material

I Talko ingår sociolingvistiska intervjuer med svenskspråkiga från olika delar av Finland. Den största delen av materialet består av inspelningar som gjordes inom projektet *Spara det finlandssvenska talet* (2005–2008). *Spara talet* var ett insamlingsprojekt med fokus på vardagligt tal. Cirka 1 000 talare, både män och kvinnor, från två åldersgrupper (20–30 år och 55–75 år) intervjuades enskilt eller parvis. Även inspelningsassistenterna hade anknytning till orterna. Av de 40–60 minuter långa intervjuerna har 20 minuter långa avsnitt ur en del av inspelningarna valts ut, transkriberats och annoterats för korpusen.

I Talko ingår också 29 kortare inspelningar på 3–5 minuter ur boken *Från Pyttis till Nedervetil* som utgavs av Gunilla Harling-Kranck 1998. Inspelningarna är gjorda 1959–1987 och de flesta informanter är födda mellan åren 1880–1905.

Material i Talko 3.0

	Inspelningar	Informanter	Orter	Token	Sekunder
Nyland	105	121	26	373333	118 705
Språköarna Tammerfors och Kotka	9	11	2	30404	11 079
Åboland	56	62	10	198400	63 519
Åland	35	41	16	117217	35 910
Österbotten	146	156	37	554103	169 799
totalt	**351**	**391**	**91**	**1273457**	**399 012**

Versioner

Talko 3.0: Från 13.01.2020– Talko 3.0 innehåller 351 inspelningar, varav 322 inspelningar från *Spara talet* och 29 *Från Pyttis till Nedervetil*.

Talko 2.1: Från 3.4.2018–18.11.2020. Talko 2.1 innehöll 281 inspelningar, varav 252 från *Spara talet* och 29 *Från Pyttis till Nedervetil*. Ny version av användargränssnittet Glossa.

Talko 2.0: 9.3.2017–3.4.2018. Talko 2.0 innehöll 271 inspelningar, varav 243 från *Spara talet* och 28 *Från Pyttis till Nedervetil*.

Talko 1.0: Juni 2015–8.3.2017. Talko 1.0 innehöll 186 inspelningar, varav 158 från *Spara talet* och 28 *Från Pyttis till Nedervetil*.

Talko 0.1: Augusti 2014–juni 2015. Talko 0.1 innehöll 100 inspelningar från *Spara talet*.

Licens

Första gången användaren loggar in i Talko måste hen godkänna den licens under vilken materialet publicerats:

CLARIN ACA (Academic) End-User License +NC +LOC +PRIV +ND 1.0

Licensen i sin helhet finns som bilaga till detta dokument men några centrala aspekter är följande:

Korpusen kan användas för undervisning och forskning. Korpusen består av ljudinspelningar som innehåller personuppgifter som omfattas av EU:s dataskyddsförordning.

SLS förutsätter även att du som användare följer de etiska principer som fastställts av Forskningsetiska delegationen.

Transkriptioner

I korpusen ingår två typer av utskrifter: en ljudenlig utskrift och en ortografisk utskrift. Utskrifterna följer det som sägs på inspelningarna ord för ord. Var uppmärksam på att en utskrift alltid är en tolkning av inspelningen. Transkriptionerna har gjorts av 22 olika personer och arbetet pågick under närmare 10 års tid. Riktlinjerna för transkriberingen har till viss del modifierats under arbetets gång. Det har medfört att det ibland finns inkonsekvenser. Det är därför viktigt att du som använder korpusen själv lyssnar på ljudfilerna och gör din egen bedömning av materialet.

Ljudenlig transkription

I den ljudenliga utskriften används en grov fonetisk transkription med det svenska alfabetets tecken. Långa vokaler markeras med kolon och långa konsonanter dubbeltecknas. *Sje*-ljudet anges med *sj*. *Tje*-ljud med hörbart t-förslag anges med *tj*. *Ng*-ljudet markeras med *ng* och om *g*

hörs med *ngg*. Därtill markeras supradentalt uttal av *rs* med *ssj* eller *rsj* men det finns en viss inkonsekvens i utskrifterna. Övriga supradentaler markeras inte utan skrivs ut som konsonantkombinationer (*rt*, *rn*). Pauser anges med punkt inom parentes. Antalet punkter signalerar pausens längd, så att fler punkter avser en längre paus: (.), (..) eller (...).

Närmare information om den ljudenliga transkriptionen finns i manualen för transkriberare på www.sls.fi/talko.

Ortografisk transkription

Den ortografiska utskriften följer standardsvensk stavning enligt *Svenska Akademiens Ordlista* (SAOL). För ord som inte ingår i SAOL används *Ordbok över Finlands svenska folkmål* (FO) och *Finlandssvensk ordbok* (FSOB) som stöd vid skapandet av den ortografiska formen.

En del förenklingar görs. Till exempel för verbens del förenhetligas böjningsformer enligt standardsvenskt mönster: preteritumformerna *lest*, *le:st*, *le:ste*, *las* blir alla *läste* i den ortografiska utskriften, supinumformerna *lesi*, *lest, le:st* blir *läst*. Ibland är kontexten avgörande för den ortografiska formen. Verb av första konjugationen kan ha samma uttal för alla böjningsformerl: uttalet *ta:la* kan förekomma som infinitiv, presens, preteritum och supinum och kontexten avgör vilken form som väljs i den ortografiska utskriften. Det innebär att även i den ortografiska transkriptionen ingår ett moment av tolkning.

I vissa fall har flertal olika uppslagsformer som ingår i FO sammanförts till en enda i den ortografiska transkriptionen, t.ex. adverbet *här*, som används för bland annat uttalen *hä:r*, *he:r*, *jä:r*, *sjenn*, *hije:*, *ije:nan*. På motsvarande sätt har pronomenet *he* sammanförts med *det*.

Närmare information om den ortografiska transkriptionen finns i manualen för transkriberare. Manualen finns på www.sls.fi/talko.

X- och g-tagg

Ord och former som inte finns i SAOL markeras med taggen x. Det gäller t.ex. dialektala ord, ord från andra språk och slangformer av namn.

En del av de x-taggade orden har försetts med direktlänkar till *Ordbok över Finlands svenska folkmål* (FO) och *Finlandssvensk ordbok* (FSOB) så att användaren vid behov kan få mer information om ordet. Länkningen har gjorts för de mest frekventa orden men är inte komplett.

Grammatiska former som inte förekommer i standardsvenska markeras med taggen g. I Talko 3.0 är det enbart de sk. substantiverade passiva infinitiven på -as(e/i), *sji:dase* 'skidandet', *programme:ras* 'programmerande', och *te:ve:skådas* 'tevetittande' som markerats med g-tagg.

Annotering

Materialet i Talko har försetts med grammatisk information (ordklass och morfologisk analys) i form av de taggar som används i Stockholm-Umeå-korpusen (SUC). Dessutom har varje ord

försetts med lemma (dvs. grundform). Annoteringen har gjorts automatiskt för de allra flesta av inspelningarna med hjälp av en statistisk taggare.

Taggningsresultatet har kunnat förbättras genom att manuellt annoterat material använts som träningsmaterial för den statistiska taggaren. I Talko 3.0 ingår 15 Spara talet-filer och 5 Från Pyttis till Nedervetil-filer som annoterats manuellt.

Korrektheten är i medeltal 93,93 procent men kan variera mellan olika dialektområden. De fel som uppstår vid den automatiska taggningen är inte slumpmässiga utan vissa ordklasser blandas oftare ihop än andra. Vanliga fel är t.ex. att adjektiv taggas som adverb och tvärtom, och att egennamn blir taggade som substantiv.

Du kan läsa mer om korpusen och annoteringen i artiklarna Talko – korpus över den talade svenskan i Finland: Korpusbygge i teori och praktik (Södergård och Leinonen 2017) och Ordklasstaggning av finlandssvenskt talspråk (Leinonen 2015).

Hänvisa till Talko

Hänvisa till Talko genom att ange vilken version du har använt för dina sökningar:

Talko – korpus över den talade svenskan i Finland. V. 3.0. Svenska litteratursällskapet i Finland. www.sls.fi/talko

Kontakt

Om du har frågor eller stöter på problem när du använder Talko får du gärna kontakta SLS arkiv, arkivet@sls.fi.

Tack!

SLS arkiv tackar alla som har varit med och gjort Talko möjligt. Förverkligandet av korpusen hade inte varit möjligt utan alla de timanställda som jobbat med att göra utskrifter. Ett ovärderligt jobb gjorde naturligtvis även alla inspelningsassistenter, över 60 personer, som gjorde intervjuer för Spara det finlandssvenska talet. Ett stort tack även till alla personer som lät sig intervjuas för projektet!

Personal

Lisa Södergård
Therese Leinonen
Katja Koskinen
Sara Rönnqvist
Janina Öhman
Ann-Sofie Grönroos

Tekstlaboratoriet

Korpusen har möjliggjorts genom ett samarbete med Tekstlaboratoriet , som är en del av Institutt for lingvistiske og nordiske studier vid universitetet i Oslo.

Bilaga: Licens för talspråkskorpusen Talko
CLARIN ACA (Academic) End-User License +NC +LOC +PRIV +ND 1.0

Resource: This license applies to the resource that points to this license document.

Copyright holder: Details about the copyright holder(s) are provided in the documentation of the Resource.

The Copyright holder grants the End-User a free, non-exclusive and perpetual (for the duration of the copyright) right to use and make copies of the Resource for educational, teaching or research purposes as such, as modified, or as part of a compilation or derived work. The permission applies to all known or future modes and means of communication and includes a right to make modifications enabling the use of the Resource on other devices and in other formats. Distribution of copies is not allowed.

Additional license terms as defined in the Terms of Service Agreement:

Identification and Access conditions

ID: The user needs to be authenticated or identified.

General Use conditions

BY: Attribution, i.e. acknowledgement of authorship, is required.

NC: The content is available for non-commercial purposes only.

LOC: The content is available only at a single location, center, or site.

PRIV: There are personal data in the resource.

Distribution conditions

NORED: The user is not permitted to redistribute the resource.

ND: The user is not permitted to make derivative works, i.e. works containing copyrighted parts of the original.

Other: There are other non-standard conditions in the license that the user should pay attention to:

The corpus consists of audio recordings containing personal data protected by the General Data Protection Regulation.

This license has been made in compliance with copyright agreements by WIPO – the World Intellectual Property Organization. The rights granted in this license shall be so interpreted that in case applicable intellectual property laws grant rights not mentioned in this license, they are also regarded as part of the rights to be licensed; the purpose of this license is not to restrict any rights intended to be licensed within different legal systems. Additional rights to the Resource may be agreed separately in writing.