# Donera prat för AI-utveckling till Språkbanken i Finland

*Krister Lindén, Mietta Lennes, Tommi Jauhiainen, Mikko Kurimo, Aleksi Rossi*

*Det behövs tusentals timmar med vardagsprat för forskning och innovation i samhällsvetenskap, humaniora och artificiell intelligens. För att utveckla nättjänster som förstår svenska och finska i Finland har Språkbanken i Finland tillsammans med Yle, det nationella public service-bolaget för radio, tv och webb, samlat in 4 000 timmar vardagsfinska via nätet. Insamlingen genomfördes med understöd från statens utvecklingsbolag Vake. En insamling av vardagsprat på finlandssvenska är på gång i samarbete med Svenska litteratursällskapet i Finland, och målet är ett liknande projekt även för samiska i Norden.*

## 1. Inledning

Det finns redan kommersiella system som använder AI med taligenkänning, såsom Apples Siri och Googles Alexa. Men många fler idéer väntar på ett kommersiellt genombrott, något som i viss mån beror på bristen på lämpliga språkresurser. En stor, öppet tillgänglig språkresurs möjliggör och påskyndar utvecklingen av språkbaserade AI-komponenter och applikationer. Öppet tillgängliga komponenter sänker tröskeln för att pröva nya idéer, samtidigt som man kan senarelägga beslutet om vilken kommersiell teknik som ska användas i produktionsfasen. Det pågår en världsomspännande insats av Mozilla Common Voice (https://commonvoice.mozilla.org/), men det initiativet har som mål att samla in uppläst tal. Färdiga manus tenderar att få människor att använda standardiserat icke-vardagligt tal.

I det projekt vi beskriver var vårt mål specifikt att samla vardagligt, spontant tal från ett stort antal talare. Vi redogör här för den process som ledde till att Vake, det finländska statliga utvecklingsbolaget (numera Ilmastorahasto Oy), beslutade finansiera utvecklingen av en insamlingsplattform för kampanjen *Lahjoita puhetta* för insamlingen av finskt talspråk; varför Språkbanken i Finland (https://www.kielipankki.fi/sprakbanken/) valdes som dataförmedlare; hur Språkbanken har förberett sig på att göra stora personuppgiftssamlingar tillgängliga och hur det nationella public service-mediebolaget Yle utformade mediekampanjen för att få så många som möjligt att donera prat.

## 2. Applikationer för specialbehov

Att söka i talinspelningar efter innehåll är felbenäget, även om det finns tekniker för att hitta enstaka ord och fraser i talsegment. Ett annat tillvägagångssätt är att konvertera tal till textavskrifter och använda befintliga verktyg för textanalys. Man vill kanske till exempel räkna hur många inspelade telefonsamtal som nämner vissa stickord i en robocall-undersökning. Exempel på en mer komplex användning är analyser av samtal för att övervaka telefonförsäljning. En annan tillämpning är automatisk translitterering av intervjuer som journalister eller forskare har gjort. Ifall man snabbt kunde hitta ett citat från talsignalen skulle det avsevärt påskynda verifieringen av detaljer i sådana intervjuer. En bättre sökbarhet för talinspelningar ökar också användbarheten av videoinspelade debatter för verifiering i efterhand, exempelvis för debatterna i samband med beslut i riksdagens plenum.

Automatisk taligenkänning används för traditionell diktering, när man behöver skriva meddelanden i situationer där händer och ögon har andra uppgifter. Diktering anpassad till en specifik persons tal fungerar redan någorlunda bra, till exempel på mobila enheter. Det gäller särskilt när bakgrundsljuden är få eller den som talar befinner sig nära mikrofonen.

Med förbättrad talbehandling kan tv-program och föreläsningar textas automatiskt, antingen direkt från originalljudet eller med återdiktering. Specifika grupper, som hörselskadade, skulle ha stor nytta av automatisk textning i realtid. Tillförlitligt fungerande, genreoberoende textning av talspråk är en god grund för automatisk översättning och tolkning, något som har otaliga användningsområden i en globaliserad värld.

Det moderna samhället kräver många digitala användarkompetenser, bland dem färdigheten att använda mobila enheter. Någon med nedsatt syn eller vars fingerfärdighet inte räcker till för en viss enhet, kan därför bli utestängd från många tjänster. Ofta kan kraven som ställs på användarna avhjälpas med ett röstaktiverat användargränssnitt för tjänster på deras modersmål. Intelligenta tillämpningar kan komplettera eller till och med ersätta personlig service för åldringar och funktionshindrade och erbjuda dem möjligheter till ett självständigt liv och en förbättrad livskvalitet. Å andra sidan, om det finns ett röstgränssnitt som fungerar dåligt, kan det väcka misstro och leda till att användarna börjar undvika tjänsten. Inom sjukvården kan brister i användargränssnittet rentav utgöra en säkerhetsrisk.

I språkinlärningsapplikationer är talgränssnitt anpassade för specifika användare mer användbara. Interaktion och muntliga färdigheter betonas i dagens samhälle och yrkesliv och blir en allt viktigare del av språkinlärningen. Invandrare i Finland har en stor fördel av goda muntliga kunskaper i finska,

både på arbetsmarknaden och för att bygga upp sina sociala nätverk. En stor databas med transkriberat vardagstal är en bra referens. Men det behövs också andra typer av data för att tillförlitligt mäta uttalsegenskaper i enskilda språkinlärares tal och för att modellera deras tal och kommunikation i en verklig interaktion.

Det finns också användningsfall där talet som ska analyseras inte behöver presenteras i textform, utan där analysen härleds direkt från talet. Sådana är till exempel automatisk talaridentifiering eller automatiserad analys av en användares ålder, vakenhetstillstånd eller hälsa. De funktionerna är användbara vid anpassningen av olika applikationer och tjänster, också fastän noggrannheten är lägre än 100 procent. Även när applikationer inte kräver att talet presenteras i textform behövs det ändå stora träningskorpusar med taldata annoterade med personliga och hälsorelaterade egenskaper.

## 3. Behovet av talkorpusar

I början av 2000-talet var den finländska talteknikens och talspråksforskningens resurser spridda runtom i Finland i relativt små team. USIX – *Uusi käyttäjäkeskeinen tietotekniikka* [sv. Det nya användarcentrerade teknikprogrammet för informationsteknik] lanserades 1999 och finansierades av Tekes (numera Business Finland). Programmet resulterade i nya projekt och samarbete mellan forskargrupper och gynnade forskningen inom finländsk tal- och språkteknologi. Med medel från undervisningsministeriet genomfördes en undersökning av talspråksforskningen i Finland (Toivanen och Miettinen 2001). Ett av de viktigaste resultaten var insikten om att det krävdes investeringar i tillgången på digitala taldata för att främja utvecklingen av forskning och teknologi för att kunna bearbeta talad finska.

Tillgången på taldata är en förutsättning för forskning inom talat språk och utveckling av taltekniska applikationer, inklusive talgränssnitt. Syftet med konsortieprojektet *Integrerade resurser för talteknologi och talspråkforskning i Finland* (SA-Puhe), som finansierades av Finlands Akademi 2003–2004, var att ta itu med behovet av allmänna riktlinjer och metoder för forskare att i samverkan samla in, kommentera och dela talkorpus. Under projektet samarbetade fonetiker och språkforskare vid Helsingfors universitet med Laboratoriet för akustik och signalbehandling vid Helsingfors tekniska högskola och med CSC–IT Center for Science.

SA-Puhe-projektet tog tag i behovet av en centraliserad infrastruktur för lagring, delning och underhåll av både taldata och relaterade kommentarer för forskningsändamål. Plattformen skulle byggas på ett objektorienterat databassystem som hade utvecklats vid Tekniska högskolan i Helsingfors inklusive

ytterligare ett samarbete med Helsingfors universitet under 1990-talet (Karjalainen och Altosaar 1993; Altosaar, Millar och Vainio 1999). Databassystemet skulle ha ett grafiskt gränssnitt (Altosaar och Lennes 2005). För att göra det möjligt för forskare att bidra med, dela och underhålla sina utskrifter och strukturerade kommentarer till talinspelningarna, utvecklades en första version av ett talannoteringsprogram (Puh-Editor) vid CSC–IT Center for Science (Grönroos och Miettinen 2004).

Tyvärr var det inte möjligt att slutföra integrationen av komponenterna i taldatabasplattformen och editorn under finansieringsperioden. Under projektets gång togs det dock fram allmänna riktlinjer för talannotering med hjälp av språkforskarna (Lennes och Ahjoniemi 2005). Dessa riktlinjer visade sig vara användbara när idén om big data för talbearbetning återupplivades, inspirerad av framstegen inom talteknologi med neurala nätverksteknologier.

Processen som ledde fram till kampanjen *Lahjoita puhetta* inleddes med möten mellan en ad-hoc-grupp av företag och offentliga organisationer år 2018. Statens utvecklingsbolag Vake beställde våren 2019 en förstudie över behovet av finska språkresurser för artificiell intelligens av FIN-CLARIN och Språkbanken i Finland (Kielipankki). Målet var att specificera vilka insatser som krävdes för att möjliggöra en bred användbarhet av de språk som talas i Finland i olika AI-tillämpningar. Satsningen inleddes med finska, som det mest talade språket i Finland. Språkbanken intervjuade över 50 kommersiella och offentliga organisationer i Finland. Ett av de åtta utvecklingsmål som identifierades var en stor korpus av spontant tal, precis som konstaterats i förstudien publicerad i oktober 2019.

Språkbanken i Finland samarbetade med Finlands nationella mediebolag (Yle) och Statens utvecklingsbolag (Vake Oy) i *Lahjoita puhetta*. Experter från Helsingfors universitet, Aalto-universitetet och Åbo universitet deltog också i projektet. Vake utarbetade dataskyddsanalysen och de juridiska dokumenten med hjälp av företaget 1001 Lakes Oy, och juridiska rådgivare från Helsingfors universitet och Yle bidrog med att utarbeta det juridiska ramverket för insamlingskampanjen.

## 4. FIN-CLARIN och Språkbanken i Finland

Sedan 2009 har FIN-CLARIN funnits på den nationella kartan över forskningsinfrastrukturer som upprätthålls av Finlands Akademi. FIN-CLARIN-konsortiet består av alla finska universitet som bedriver språk- och språkteknologisk forskning, Institutet för de inhemska språken (Kotus) samt CSC–IT Centre for Science. FIN-CLARIN upprätthåller Språkbanken i Finland, genom vilken medlemmarna i konsortiet erbjuder olika språkresurser som korpusar, lexikala resurser och verktyg.

Ända sedan Språkbankens tillkomst år 1996 har målet varit att både korpusar och verktyg ska vara tillgängliga för forskarvärlden på ett så effektivt sätt som möjligt. Eftersom det har ägnats liten uppmärksamhet åt att göra material och verktyg tillgängliga för företag, licensieras många språkresurser med en icke-kommersiell begränsning. I många fall har också upphovsrätt eller dataskydd lett till begränsade licenser. I FIN-CLARIN ansvarar CSC för det tekniska underhållet och Helsingfors universitet för anskaffningen och kurateringen av korpus och verktyg.

## 5. Taldata för kommersiella ändamål

Att transkribera tal till text är en subjektiv process. En utskrift produceras för ett visst syfte och den återspeglar de val som gjorts av en enskild transkriberare. Oavsett transkriptionssystem kan en transkription inte spegla alla relevanta egenskaper för naturlig interaktion och nyanser i talet. Dessa inkluderar tillfälliga variationer i talljud eller andra ljud, liksom grundläggande prosodiska egenskaper: röstkvalitet, tonhöjd, intensitet, talhastighet och pauser. De här egenskaperna bidrar inte bara till intrycken av melodi, accenter och rytm, utan också till de upplevda betydelser, avsikter och attityder vi hör och förstår i varandras tal samt till gester, uttryck, blickar och andra handlingar relaterade till interaktionssituationen och dess sammanhang. Det primära målet för transkriptionen av det insamlade talet är att tillhandahålla en fonematiskt korrekt transkription av ljuden i signalen, som senare konverteras till standardiserat tal för sökbarhet och möjliggör vidare forskning och utveckling i automatiserad språkbearbetning.

Konstruktionen av säkra, persondatavänliga röstanvändargränssnitt kräver i vissa fall att komponenterna i en applikation kan användas utan överföring av persondata från en tjänst till en annan, till en tredje part eller till en annan stat. Det här talar för att talbearbetningens komponenter bör vara öppet tillgängliga med öppen källkod.

Talkorpusar som distribueras av Språkbanken i Finland, till exempel från Finlands riksdagsplenum, som innehåller inspelningar av riksdagens plenarsammanträden från 2008 till 2020, samt utskrifter av dessa, är licensierade som CC-BY-NC-ND. NC är en förkortning för icke-kommersiell. Den innebär att materialet inte får användas för kommersiella ändamål. Ett sätt att utöka det kommersiellt användbara talmaterialet är att omförhandla licenser för korpusar för att tillåta en sådan användning. Även om det i det nämnda fallet med parlamentets plenarsessioner fortfarande är möjligt, går det ofta inte att omförhandla tillgången till talmaterial efter att det har samlats in. Av den anledningen är det viktigt att se till att nytt talmaterial samlas in på ett sätt som möjliggör kommersiell användning.

## 6. Juridiska aspekter på insamling av talmaterial

Det juridiska ramverket i EU ska tillhandahålla ett interoperabelt utrymme för olika slag av verksamhet. Medan det juridiska ramverket harmoniserar mycket av verksamheten i samhället i övrigt, har forskningspraxis ofta överlåtits till nationella överväganden. Det här påverkar möjligheterna till utbyte av data genom en forskningsinfrastruktur som CLARIN, eftersom vi behöver en gemensam rättslig grund som går att tillämpa på forskning i alla EU-länder. Dessutom är forskningen inte enbart begränsad till den akademiska världen. För att fördela resurserna inom ett land, behövs det därför ofta lösningar som också omfattar industriell och kommersiell forskning.

De immateriella rättigheterna i det juridiska ramverket inom EU har diskuterats ingående (Kelli & al. 2016, 2018b, 2019a) av medlemmar i CLARIN Legal Issues Committee. CLARIN rekommenderar Creative Common-licenser när det är möjligt (Oksanen & al. 2011). För alla datamängder, inklusive dem som inte kan göras öppet tillgängliga, erbjuder CLARIN ett klassificeringssystem för licensmetadata (Oksanen & al. 2010) för att informera användaren om potentiella begränsningar när man får tillgång till en datasamling. För datasamlingar som inte kan göras öppet och offentligt tillgängliga erbjuder CLARIN standardlicensmallar för deponering av data som ska delas via CLARIN (Kelli & al. 2018a). De immateriella rättigheter som är relevanta för att dela forskningsdata har granskats tämligen ingående av CLARIN under de senaste tio åren, men nya möjligheter öppnar sig via EU:s text- och datautvinningsdirektiv (Kelli & al. 2020b).

Under de senaste åren har konsekvenserna av EU:s allmänna dataskyddsförordning (GDPR) blivit allmänt kända (Kelli & al. 2021). Enskilda EU-medlemsländer har ett visst utrymme att göra undantag för forskning, vilket har lett till varierande praxis när det gäller delning av personuppgifter för akademiska forskningsändamål (Kelli & al. 2019b, Lindén & al. 2020). Resurser som innehåller personuppgifter är bland de resurser som inte kan göras tillgängliga utan skyddsåtgärder, och CLARIN uppdaterar nu licensmallar för att återspegla hur personuppgifter fortfarande kan delas säkert och kontrollerat för akademisk forskning (Kelli & al. 2020a). Trots att alla data inte kan göras öppet tillgängliga är det möjligt att använda data som man har laglig tillgång till för att skapa öppet tillgängliga språkmodeller (Kelli & al. 2020c).

Från början stod det klart att behandlingen av personuppgifter i *Lahjoita puhetta*-projektet måste bedrivas både juridiskt och etiskt hållbart. Alla centrala aktörer i *Lahjoita puhetta*, Språkbanken vid Helsingfors universitet, Vake och Yle är offentliga organisationer som inte kan bortse från sådana aspekter. För att förstå de eventuella problem behandlingen av personuppgifter kunde

medföra för individer gjordes därför en noggrann riskbedömning. Dataskyddsmyndigheten förordar ett balanstest för den som vill yrka på legitimt intresse som laglig grund för insamling. Efter att ha gått igenom de sex stegen i testet, stod det klart att insamlingen uppfyllde kraven och att intressena i forskningen inte åsidosattes vare sig av donatorernas intressen eller deras grundläggande rättigheter och friheter. Det gjordes också en konsekvensbedömning av dataskyddet med tanke på möjliga risker i anslutning till behandlingen av personuppgifter, där särskilt den omfattande bearbetningen av materialet och utvecklandet av ny teknologi relaterades till ändamålet. Språkbanken i Finland kommer att ge tillgång till taldata när en tillräcklig mängd material har donerats och processen med att ge rättigheter till kommersiella aktörer är slutförd. Detta borde vara möjligt i början av år 2022.

En närmare beskrivning av de juridiska dokument som utarbetades för insamlingen av personuppgifter så att taldata kan delas inom EU, finns i Lindén & al. (2022). Det faktum att uppgifterna samlades in också för industrins bruk gör att avtalsramverket är relevant även i nordiska sammanhang, eftersom den industriella användningen regleras av EU:s gemensamma dataskydd.

## 7. Hur man planerar en massiv taldatainsamling

Den första fasen av *Lahjoita puhetta*-kampanjen fokuserade på att få ihop 10 000 timmar vardagsfinska representativ för hur finländarna i dag talar i olika sammanhang.

Kampanjen hade en föregångare i ny metodik och nya sätt att erhålla taldata över internet som hette *Prosovar* och genomfördes av forskare vid Åbo universitet. Målet med *Lahjoita puhetta* var inte bara att samla in en stor mängd vardagstal, utan även att nå ut till så många som möjligt. I marknadsföringen betonades det att donationer av alla varianter av talad finska var välkomna, också från talare som inte hade finska som modersmål. Dataskyddet och instruktionerna förutsatte ändå en viss nivå på språkkunskaperna.

För att hitta en balans mellan de materiella målen, de tekniska möjligheterna och de tillgängliga resurserna hölls designworkshoppar för alla intresserade. Under dem samlade man in idéer för både insamlingskoncept och användningsområden från forskare både inom industri- och universitetsvärlden. Största delen av den praktiska planeringen överläts till personalen på Yle, som ansvarade för att mobilisera allmänheten genom sina radio- och tv-kanaler. Yle designade bilderna, filmerna och texterna i webbapplikationen och de nedladdningsbara apparna. Det utformades också tekniska mallar för design av teman med olika typer av innehåll riktade till specifika målgrupper.

Hösten 2019 och våren 2020 bestämdes potentiella användningar, målgrupper samt obligatoriska och valfria funktioner i gränssnittet på workshoppar, som organiserades av utvecklingsföretaget Solita enligt en metodik som delvis byggde på Design Thinking. Senare testades en rad nyckelfunktioner med pappersprototyper och semi-interaktiva verktyg. Förslag gjordes av professionella och erfarna webbservicedesigners, och en del funktioner testades i praktiken.

Nyckelfrågorna och utmaningarna för utformningen av användargränssnittet gällde att få personerna att tala fritt och vinna deras förtroende så att de kände sig bekväma, samtidigt som gränssnittet måste uppfylla de juridiska begränsningarna och enkelt presentera den nödvändiga informationen. Efter att ha testat några idéer bestämde sig Yle för en video, en bild och ett textinnehåll som lockade en person att prata, kombinerade med en funktionell knapp med vilken talaren kunde starta och stoppa inspelningen.

Det fördes också diskussioner kring något slag av spelelement som meddelade användarna hur mycket de hade donerat, till exempel i form av jämförbara resultat för att upprätthålla intresset eller sociala element som att dela resultat eller samla team. Slutligen inkluderades enbart den totala tid som donerats som ett sådant spelelement.

Inledningstemat *Harjoitellaan ensin* (sv. Vi börjar med att öva) startade med en testsession av inspelningen med användaren. Sessionen var samtidigt ett utmärkt tillfälle att betona att främst AI-forskare skulle använda inspelningarna och att påminna användaren om dataskyddet. Den tekniska plattformen presenterade metadatafrågor om dialektbakgrund och grundläggande demografi som åldersgrupp, kön, modersmål, nuvarande boningsort, yrkeskategori och utbildningsnivå. Även användarens tekniska plattform noterades för statistiska ändamål, men inga positionsdata samlades in.

Yle tog slutligen fram ett 40-tal okomplicerade teman för att stimulera insamlingen. Utöver övningstemat samlades nästan hälften av uppgifterna in med följande av de tolv populäraste temarubrikerna:

Rakkain eläimeni *(Mitt käraste djur)*
Mistä kodikkuus syntyy? *(Vad gör ett hem trevligt?)*
Tärkeä esineeni *(Ett viktigt/kärt ting)*
Lempivaate *(Mitt favoritplagg)*
Mikä suututtaa? *(Vad gör mig arg?)*
Turhat tavarani *(Mina onödiga prylar)*
Mitä opimme? *(Vad lärde vi oss?)*
Entisajan lemmikit *(Husdjur från förr)*

Katson ikkunasta *(Vad jag ser från mitt fönster)*
Kuva-arvoitus *(En bildgåta)*
Kerro aamiaisesta *(Berätta om din frukost)*

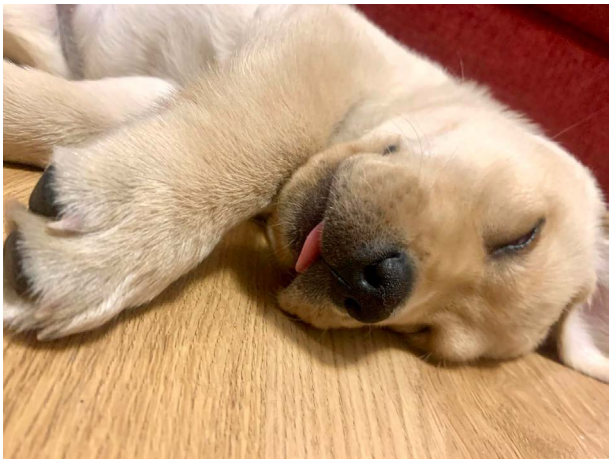

*Bild 1. Mitt käraste djur var ett populärt tema i pratdonationerna. Foto: Språkbanken*

Yle gjorde humoristiska inforeklamer med uppmaningar till allmänheten om att donera tal. Dessa sändes mellan programmen i radio- och tv-kanalerna sommaren och hösten 2020, med några repriser våren 2021.

## 8. Tekniskt genomförande

Talet till *Lahjoita puhetta* (https://lahjoitapuhetta.fi/) gick att donera via webbläsare eller mobilappar. Representanter både från industri och universitetsvärlden utvecklade de allmänna specifikationerna. Företaget Solita utvecklade apparna. Mjukvaruplattformen finns publicerad som en öppen källkod som tillåter andra organisationer att bygga egna system för talinsamlingar, specialiserade insamlingskampanjer eller för liknande kampanjer i andra länder.

Teknisk röstkvalitet är ett komplicerat ämne. Det är till exempel nödvändigt att mikrofonen finns nära användaren. Goda råd med avsikten att få talarna att slappna av genom att låta telefonen ligga på bordet hade lett till mer eko och en svagare signal. Att låta en grupp människor diskutera var också uteslutet. En fritt flödande gruppdiskussion hade visserligen haft uppenbara fördelar jämfört med en talsituation med endast en talare. Den hade däremot inneburit svåra tekniska utmaningar både med deltagarna nära varandra, med samma mikrofon, och långt borta från varandra med var sin enhet för att minimera störningar och ekon. Dessutom hade detta krävt en synkronisering av flera signaler eller

en registrering av vilka telefoner som spelade in diskussionen. Av den här anledningen genomfördes inga tester med gränssnitt och teman för dialogflöden.

Inspelningarna hölls enkla, genom att bara spela in talsignalen i högsta möjliga förlustfria format och lagra den med metadata om system, telefonmodell och version. Metadata möjliggör vissa efterbearbetningskorrigeringar med till exempel ljudutjämning per mikrofontyp. En rudimentär mätare för att ge feedback till användaren om en acceptabel signalnivå övervägdes, men den implementerades inte, bland annat för att spara på mobilernas batterier. Användartester gav oss också orsak att betvivla ändamålsenligheten av en sådan mätare. För det första skulle mätaren ha utgjort en distraktion eller med största sannolikhet ha ignorerats. För det andra skulle informationen om hur mätaren skulle tolkas ha belastat användargränssnittet och för det tredje skulle förbättringen av signalen inte ha varit väsentlig, eftersom användaren förmodligen bara hade talat närmare mikrofonen under en kort stund.

Till slut instruerades användarna helt enkelt att tala fritt i sin egen miljö. En tydlig signal i en störningsfri miljö är ofta att föredra. För tillfället har inspelningarna lite mer variation när de innehåller en del störningar, som bakgrundsljud från andra människor eller vind utomhus. Enligt användartesterna gjordes de flesta inspelningssessionerna i tysta inomhusmiljöer. En fördröjd sändning i bakgrunden av lokalt lagrade inspelningar för uppladdning till molnet förbereddes ifall användaren inte hade en stabil internetuppkoppling, men det var förmodligen inte en avgörande funktion.

Webben, Android och iOS valdes som plattformar för smartphones, surfplattor och datorer med mikrofoner. Den tillhörande webbplatsen informerade också användarna om kampanjen och Yle publicerade artiklar på sin kampanjsajt. För att väcka användarnas förtroende för kampanjen gavs apparna ut via Yle istället för via separata dedikerade eller kampanjspecifika app-konton.

Servicearkitekturen bestod av flera gränssnitt på olika plattformar, webbservicetjänster och databaser för insamling av data i molnet. Genom att fördela ansvaret mellan Yle och Helsingfors universitet kunde det juridiska ansvaret för de olika parterna begränsas, så att Yle som genomförde kampanjen fick tillgång till aggregerade användningsdata utan att få tillgång till och ansvar för persondata. Systemet utvecklades för enspråkigt bruk. Anpassningen och lokaliseringen till andra språk som finlandssvenska och samiska hölls i åtanke och testas under 2022.

För att beakta GDPR och möjliggöra radering av bidrag tillåts enkel radering av användarbidrag genom en lång slumpmässig identifierare som ges till användaren vid taldonationen. Det finns inga andra användarspecifika identifierare. Man måste ändå alltid överväga om enskilda användare kan identifieras med sina metadata ifall den deltagande gruppen är liten eller en kombination av vissa

metadata är mycket specifik. Exempelvis kan antalet män i en viss åldersgrupp på ett litet geografiskt område med en speciell dialektbakgrund potentiellt resultera i att en grupp människor skulle kunna identifieras både med den insamlade informationen och i den verkliga världen. Den tekniska plattformen som sådan begränsar inte insamlingen av specifika metadata eftersom GDPR-kompatibel behandling av personuppgifterna sker på ansvar av den som senare behandlar uppgifterna eller publicerar resultat som bygger på dessa.

Under våren 2021 deltog kampanjens Android- och iOS-mobilapplikationer i den årliga marknadsföringstävlingen *Grand One* för webbapplikationer. Applikationerna vann första pris i kategorin mobiltjänster och fick ett hedersomnämnande i kategorin för bästa dataanvändning. Yle skickade också in kampanjen till den årliga tävlingen för europeiska mediebolag *Prix Europa*, där kampanjen vann kategorin *Bästa europeiska digitala ljudprojekt 2021* för tv-, radio- och online-produkter bland 684 bidrag från 26 länder.

## 9. Det insamlade materialet

Kampanjen *Lahjoita puhetta* engagerade över 25 000 medborgare i Finland. De donerade över 240 000 talprover med totalt cirka 4 000 timmar vardagstal som kan användas både akademiskt och industriellt för att utveckla och forska i språk- och AI-applikationer. Cirka 3 500 timmar samlades in på bara ett halvår och ytterligare 500 timmar under halvåret därpå. Lanseringen på nationell tv i juni 2020 inspirerade till det största antalet bidrag, med närmare 500 timmar under de första månaderna. Kampanjen nådde nya målgrupper under hösten 2020, om än i långsammare takt. Mot slutet av kampanjen fokuserade man på att samla in dialekter med hjälp av regionalradion och ytterligare tio procent av materialet samlades in under en vecka kring julen 2020. Yle inrättade en kampanjsida för sina kampanjevenemang (https://yle.fi/aihe/lahjoita-puhetta). Kampanjen avslutades officiellt vid nyår 2020 men infoblänkare och repriser sändes fortfarande under våren 2021, vilket resulterade i ytterligare en rännil av bidrag. Närmare 90 procent av bidragen var mellan 10 sekunder och 3 minuter, med en medianlängd per inspelning på 30–60 sekunder.

Det samlades taldata från ett brett spektrum av åldersgrupper. Lite överraskande var 21–30-åriga kvinnor oberörda av det något tekniska upplägget och donerade det mesta av talet. Den minsta mängden tal donerades av mycket unga (1–10 år gamla) och mycket gamla (80 år eller äldre). Alla har inte tillhandahållit alla metadata, men bland dem som uppgett metadata fanns några intressanta observationer: Personer mellan 20 och 60 år gjorde cirka tre fjärdedelar av donationerna. Nästan 70 procent av donatorerna var kvinnor. Som väntat kom närmare hälften av donationerna från fyra regioner med de största städerna: Nyland

(inklusive Helsingfors och Esbo), Norra Österbotten (inklusive Uleåborg), Egentliga Finland (inklusive Åbo) och Birkaland (inklusive Tammerfors). Donationer gjordes ändå i alla regioner i Finland i över 50 städer. 95 procent av donatorerna var infödda finländare. De geografiska områdena hade ungefär samma mängd donationer per 100 000 invånare. En stor andel invånare med svenska och samiska som modersmål i vissa områden förklarar troligen ett par regioner med lite färre bidrag. Mer än två tredjedelar av uppgifterna donerades av studerande, pensionärer, lärare, entreprenörer, experter och sjukskötare (i fallande ordning efter bidragsgivarnas antal) och resten kom från mer än 30 andra yrkesgrupper. Cirka 62 procent hade högre utbildning och 28 procent gymnasieutbildning.

## 10. Avslutningsvis

Trots ett ambitiöst mål lyckades kampanjen *Lahjoita puhetta* samla in en omfattande resurs av finskt talspråk från ett stort antal talare på bara några månader. Donationer kom från alla regioner i Finland och de geografiska områdena hade ungefär samma mängd donationer per 100 000 invånare. Av de insamlade 4000 timmarna har 1 500 timmar translittererats.

Målet med kampanjen var inte bara att samla in en stor mängd tal, utan att nå ut till så många olika grupper och till så många individer som möjligt. Alla donatorer tillhandahöll inte fullständiga metadata, men bland dem som gav metadata noteras att personer mellan 20 och 60 år utgjorde cirka tre fjärdedelar av donationerna och att nästan 70 procent var kvinnor. Två grupper att överväga för framtida punktinsatser är tonåringar runt 11–20 och pensionärer runt 71–80. Bägge har distinkta egenskaper ur AI-utvecklingssynpunkt, talar i olika tonhöjd, med olika ordförråd, takt, pauser och andning. Företag överväger att utveckla AI-drivna äldreomsorgssystem, och specifika taldata för personer som är sängliggande kunde därför också vara användbara.

En liknande kampanj *Donera prat* för att samla in finlandssvenskt vardagsspråk med samma syfte lanserades av Yle i slutet av 2021 i samarbete med Helsingfors universitet och Svenska litteratursällskapet i Finland. Målet är att förbättra möjligheterna att utveckla talbaserade tjänster som förstår finlandssvenska.

En insamling av samiska planeras i samarbete med Yle, Norges public service-radio och -television NRK och universiteten i Helsingfors och Tromsö.

## Summary

Thousands of hours of everyday talk are needed for research and innovation in social sciences, humanities and artificial intelligence. To develop online services that understand everyday Finnish, the Language Bank of Finland together with the

national broadcasting company Yle has collected 4,000 hours of colloquial Finnish online with funding from the Finnish State Development Company Vake to develop the speech collection platform. A similar campaign is underway for Finnish-Swedish in collaboration with the The Society of Swedish Literature in Finland (SLS) and the goal is also to carry out a similar campaign for the Sami languages in the Nordic countries in cross-border cooperation with interested parties.

## Författare


Forskningsdirektör **Krister Lindén** är nationell koordinator för FIN-CLARIN och verksamhetsledare för Språkbanken i Finland. Dr **Tommi Jauhiainen** och MA **Mietta Lennes** är projektplanerare på Språkbanken i Finland vid Helsingfors universitet. Prof. **Mikko Kurimo** är professor i talteknologi vid Aalto universitet. MSc **Aleksi Rossi** är utvecklingsdirektör på det nationella mediebolaget Yle.


*Krister Lindén*

*Tommi Jauhiainen*

*Mietta Lennes*

*Mikko Kurimo*

*Aleksi Rossi*

## Källhänvisningar


Altosaar, T., B. Millar & M. Vainio, 1999: Relational vs. object-oriented models for representing speech: a comparison using ANDOSL data. I: Proceedings of EUROSPEECH'99, Budapest, Hungary, 5-9 Sep 1991, Vol. 2, pp. 915-918.

Altosaar, T. & M. Lennes, 2005: A Graphical Query Formation Compiler for Speech Database Access. I: The Second Baltic Conference on Human Language Technologies, Tallinn, Estonia, April 4-5, 2005, pp. 209-218.

Grönroos, M. & M. Miettinen, 2004: Infrastructure for Collaborative Annotation of Speech. I: LREC 2004, pp. 543-546.

Karjalainen, M., & T. Altosaar, 1993: An Object-Oriented Database for Speech Processing. I: Proceedings of Eurospeech 1993, Madrid, Spain.

Kelli, A., K. Vider & K. Lindén, 2016: The regulatory and contractual framework as an integral part of the CLARIN infrastructure.

Kelli, A., K. Lindén, K. Vider, P. Labropoulou, E. Ketzan, P. Kamocki & P. Stranák, 2018a: Implementation of an Open Science Policy in the context of management of CLARIN language resources: a need for changes? I: Selected papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017. Linköping University Electronic Press.

Kelli, A., T. Mets, K. Vider, A. Värv, L. Jonsson, K. Lindén & R. Birštonas, 2018b: Challenges of transformation of research data into open data: The perspective of social sciences and humanities. International Journal of Technology Management & Sustainable Development, 17(3), 227-251.

Kelli, A., A. Tavast, K. Lindén, K. Vider, R. Birštonas, P. Labropoulou, I. Kull, G. Tavits & A. Värv, 2019a: The extent of legal control over language data: the case of language technologies. I: Proceedings of CLARIN Annual Conference 2019. CLARIN ERIC.

Kelli, A., K. Lindén, K. Vider, P. Kamocki, R. Birštonas, S. Calamai, P. Labropoulou, M. Gavrilidou & P. Straňák, 2019b: Processing personal data without the consent of the data subject for the development and use of language resources. I: Selected papers from the CLARIN annual conference 2018, Pisa, 8-10 October 2018 (pp. 72-82). Linköping University Electronic Press.

Kelli, A., K. Lindén, K. Vider, P. Kamocki, A. Tavast, R. Birštonas, G. Tavits, M. Keskküla & P. Labropoulou, 2020a: CLARIN contractual framework for sharing language data: the perspective of personal data protection. I: Pro-

ceedings of CLARIN Annual Conference 2020. 05–07 October 2020, Online Edition. CLARIN ERIC.

Kelli, A., A. Tavast, K. Lindén, K. Vider, R. Birštonas, P. Labropoulou, I. Kull, G. Tavits, A. Värv, P. Straňák & J. Hajic, 2020b: The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies. I: Selected Papers from the CLARIN Annual Conference 2019. Linköping University Electronic Press.

Kelli, A., A. Tavast, K. Lindén, R. Bristonas, P. Labropoulou, K. Vider, I. Kull, G. Tavits, A. Värv & V. Mantrov, 2020c: Impact of Legal Status of Data on Development of Data-Intensive Products: Example of Language Technologies. Legal Science: Functions, Significance and Future in Legal Systems II.

Kelli, A., K. Lindén, K. Vider, P. Kamocki, A. Tavast, R. Birštonas, G Tavits, M. Keskküla, P. Labropoulou, I. Kull, A. Värv, M. Erikson, A. Vutt & S. Calamai, 2021: Sharing is caring: a legal perspective on sharing language data containing personal data and the division of liability between researchers and research organisations. I: Selected Papers from the CLARIN Annual Conference 2020. Virtual Event, 2020, 5-7 October (pp. 129-147). Linköping University Electronic Press.

Lennes, M. & S. Ahjoniemi, 2005: Puheaineiston annotaatio eli nimikointi (Version 1.01) [Annotating speech data (Version 1.0)]. Zenodo. http://doi.org/10.5281/zenodo.1205453

Lindén, K., T. Jauhiainen, M. Lennes, M. Kurimo, A. Rossi, T. Kurki & O. Pitkänen, 2022: Donate Speech – Collecting and sharing a large-scale speech database for Social Sciences, Humanities and Artificial Intelligence research and innovation. I: A. Witt & D. Fisher (eds.): The CLARIN Book. Berlin: De Gruyter.

Lindén, K., A. Kelli & A. Nousias, 2020: A CLARIN Contractual Framework for Sharing Personal Data for Scientific Research. I: Selected Papers from the CLARIN Annual Conference 2019. Linköping University Electronic Press.

Oksanen, V., K. Lindén & H. Westerlund, 2010: Laundry Symbols and License Management: Practical Considerations for the Distribution of LRs based on experiences from CLARIN. I: Proceedings of LREC 2010 Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management.

Oksanen, V. & K. Lindén, 2011: Open Content Licenses: How to choose the right one. NEALT Proceedings Series.

Toivanen, J. & M. Miettinen, 2001: Puheentutkimuksen resurssit Suomessa [Resources for speech research in Finland]. CSC – IT Centre for Science Ltd., 2001. ISBN 952-9821-76-X

## Nyckelord

infrastruktur, språkbank, korpus, talspråk, massiva datasamlingar