**Vad är språkteknologi?**

Språkteknologi handlar om hur människans språk är uppbyggt och hur en dator kan programmeras för att hantera mänskligt språk.

Språkteknologi är ett centralt område för vår språkanvändning och utvecklingen av svenskan. Språktekniska produkter genomsyrar stora delar av samhället. Det handlar bland annat om dikteringsprogram, datoruppläst e-post, söktjänster på nätet, telefontaltjänster och automatisk textsammanfattning.

Enklare och bättre kommunikation

Målet med språkteknologi är att förenkla och förbättra kommunikationen mellan människor och mellan människor och datorer. Det kan handla om att du tack vare en automatiskt uppläst översättning kan prata svenska i telefonen med någon som pratar arabiska, att du kan prata in en text i datorn i stället för att använda tangentbordet eller att din text blir automatiskt språkrättad av datorn.

Exempel på språkteknologiska områden

* Maskinöversättning
* Informationshantering
* Skrivverktyg
* Elektronisk språkinlärning
* Dialogsystem
* Taligenkänning
* Talsyntes
* Textanalys
* Elektroniska lexikon
* Tal- och textkorpusar
* Teknik och språkresurser
* Teknik och språkresurser

Språkteknologi utgörs av både tekniker och språkresurser (lexikon, korpusar etcetera). Teknikerna tar form som programvara, medan resurserna är den information som programvaran utnyttjar.

**Verktyg från Språkbanken Sam och Språkbanken Tal**

Språkbanken arbetar aktivt för att utveckla en språkteknologisk infrastruktur. Vi skapar, harmoniserar och standardiserar fria språkliga resurser och tar fram verktyg för att utforska resurserna och göra dem tillgängliga. Här visar vi några exempel från Språkbanken Sam och Språkbanken Tal.

**Edyson. Ett verktyg som gör det möjligt att snabbt få en överblick av stora ljudmaterial utan att behöva lyssna igenom hundratals timmar eller träna modeller specifikt för uppgiften. Edyson klipper upp materialet i korta ljudsegment och bildar sedan kluster med datapunkter utifrån hur de låter. På så sätt går det att skilja ut tal från applåder, hitta pauser, perioder av tystnad, brus, musik, och många andra ljudtyper. Edyson har tagits fram av Språkbanken Tal.**

**Braxen. Uttalslexikonet Braxen har utvecklades under nästan två decennier av Myndigheten för tillgängliga medier, MTM för internt bruk. Tillsammans med Språkbanken Tal har MTM nu släppt det fritt för nedladdning av forskare, talsyntesföretag, förlag och andra. Lexikonet utökas kontinuerligt med nya nyhetsord och facktermer. Det kan användas för talforskning och för att hjälpa syntetiska röster att uttala saker rätt, till exempel latinska termer och utländska namn. I lexikonet finns idag cirka 850 000 ord och namn.**

**Swedia 2000. Mer än 100 talare av olika dialekter spelades mellan 1998 och 2000 in i Swedia 2000-projektet, som var ett omfattande samarbete mellan flera universitet. Ljudfilerna innehöll persondata som var avsedda att klippas bort och tillvaratas i efter processning, men detta skedde endast i vissa fall då projektet tog slut. Materialet kom till Språkbanken Tal 2024, och personuppgifterna är nu identifierade och borttagna, så att materialet kan göras tillgängligt för forskning både hos Språkbanken Tal och Språkbanken Sam.**

**TubeN. Ett lättviktsverktyg för talforskare oclh studenter som beräknar formantfrekvenser givet en beskrivning av en munhåla, och genererar statiska vokaler från detta. Verktyget implementerar de algoritmer Liljencrantz och Fant beskrev 1975, och kan användas både till att testa hypoteser och till att ge stöd i träning och undervisning. Verktyget är utvecklat av Språkbanken Tal.**

**Fonaydyn. Ett verktyg för att undersöka fonation samt för att utforska och validera så kallad röstmappning. Rösten betraktas här mer som en sammansatt helhet än på det traditionella reduktionistiska viset, så att kopplingar mellan till exempel grundton och styrka kan fångas. Verktyget utvecklas hos Språkbanken Tal.**

**Folke.** En digital arkivtjänst med ett omfattande folkminnesmaterial som belyser människors liv och arbete, vardag och fest, erfarenheter och minnen från slutet av 1800-talet och framåt. På Folke kan alla som vill bidra till att göra samlingarna mer tillgängliga genom att transkribera handskrivna texter. Folkminnesmaterialet har samlats in av Institutet för språk och folkminnen, Isof, och digitaliserats av Språkbanken Sam.

**Lexin**. Ett nätbaserat lexikon på nitton minoritetsspråk samt svenska. Lexin är speciellt anpassat för att användas i undervisningen i svenska som andraspråk. Lexin innehåller även bilder, animationer och dialoger i videoformat. Lexikonet har tagits fram av av Institutet för språk och folkminnen, Isof, och digitaliserats av Språkbanken Sam.

**Ordregn**. Ordregn är en utveckling av det klassiska ordmolnet men placerar till skillnad från ordmolnet ord som har en liknande betydelse nära varandra och ger därför en överblick av vilka kategorier av ord som är viktiga i texten. Ett ordregn visar också mer exakt hur viktiga olika ord är. Ordregn har tagits fram av Språkbanken Sam.

**Topics2Themes**. Ett exempel på topic modelling, en språkteknologisk metod som gör det enklare att hitta, sortera och analysera stora mängder text. Verktyget har tagits fram av Språkbanken Sam.

**Projekt och samarbeten**

Språkbanken driver egna projekt och samarbetar med andra aktörer för att utveckla språkteknologi. Här visar vi några exempel.

**KB-Whisper.** KB-Whisper är en tal-till-textmodell som har tränats på 50 000 timmar tal för att lära sig omvandla talat språk till text. KB-Whisper har tagits fram av Kungliga Biblioteket som bland annat har tränat AI-modellen på äldre dialektinspelningar från Institutet för språk och fornminnen som digitaliserats av Språkbanken Sam.

**Mormor Karl är 27 år.** Projektet Mormor Karl är 27 år syftar till att skapa språkteknologiska algoritmer som kan upptäcka personuppgifter och känslig information i stora textmassor och automatiskt ersätta detta med pseudonymer. På så sätt kan personuppgifter skyddas och alla texter användas i olika slags forskning.

**Ryktesminering.** ProjektetRyktesminering, syftar till att öka förståelsen för attityder och värderingar till vaccin och undersöka hur rykten om vaccin etableras och sprids på sociala medier. Projektet är ett samarbete med Lunds universitet.

**Språkteknologi för nationella minoritetsspråk**. Språkbanken Sam arbetar med att utveckla språkteknologi som rättstavningsprogram och tangentbord för de nationella minoritetsspråken. Arbetet görs i samarbete med Giellatekno vid Universetet i Tromsö.

**SWENER-1800**. Ett digitalt textmaterial med en halv miljon ord från 1700-talet till år 1900. Materialet kan användas för att träna datorprogram att känna igen historiska namn på personer och platser i en text och för att utvärdera hur väl språkmodeller hanterar historisk svenska. SWENER-1800 är ett samarbete inom Språkbanken CLARIN mellan Uppsala universitet, Riksarkivet och Språkbanken Text.

**Svensk diakronisk korpus**. En digital textsamling med cirka 16 miljarder ord som sträcker sig från 1200-talet till nutid. Korpusen ger möjlighet att göra storskaliga studier av svenska språkets utveckling genom historien och kan också användas för språkhistoriska jämförelser mellan svenska och andra språk. Svensk diakronisk korpus är ett samarbete inom Språkbanken CLARIN mellan Uppsala universitet och Språkbanken Text.

**SuperLim 2.0**. En plattform för att testa och analysera svenska språkmodeller. Projektet är ett samarbete mellan Språkbanken Text, KB-labb, forskningsinstitutet RISE och AI Sweden

/Fyra avdelningar/

**Vi ger forskare nya möjligheter med hjälp av språkteknologi**

Språkbanken skapar möjligheter att forska i digitala text- och talmaterial med hjälp av verktyg och metoder i gränslandet mellan språkteknologi och AI. Språkbanken består av fyra avdelningar: Språkbanken CLARIN, Språkbanken Sam, Språkbanken Tal och Språkbanken Text.

Språkbanken CLARIN

Språkbanken CLARIN är en nod i CLARIN – Common Language Resources and Technology Infrastructure. Språkbanken CLARIN koordineras av Uppsala universitet, och vänder sig till forskare och andra intresserade av digitala metoder och material inom humaniora och samhällsvetenskap.

Språkbanken Sam.

Språkbanken Sam arbetar för att göra Isofs språkliga material mer tillgängligt för forskare och allmänhet med hjälp av verktyg och metoder från språkteknologi och digital humaniora. Plats: Institutet för språk och folkminnen, Isof. Göteborg, Stockholm och Uppsala.

Språkbanken Tal

Språkbanken Tal utvecklar, förvaltar och distribuerar talteknologiska resurser för svensk talforskning och svensk talteknologisk forskning. Språkbanken Tal är en forskningsenhet vid KTH.

Språkbanken Text

Språkbanken Text är en forskningsinfrastruktur för språkliga data och en språkteknologisk forskningsenhet. Vi utvecklar, förädlar och tillgängliggör fria språkliga forskningsdata enligt FAIR-principerna, med ett särskilt fokus på svenska språket genom tiderna.

Plats: Institutionen för svenska, flerspråkighet och språkteknologi, Göteborgs universitet.