Lars Borin är nationell koordinator för SWE-CLARIN som ska hjälpa forskare hantera de väldiga textmaterial som finns på nätet.

Verktyg för miljarder ord

Vad kan man göra med Språkbankens 11 miljarder ord? Eller med Kungliga bibliotekets digitalisering av samtliga svenska dagstidningar?

Ja, på egen hand går det knappast att hantera sådana enorma mängder text.

– Det är därför SWE-CLARIN finns, förklarar professor Lars Borin. Han är nationell koordinator av infrastrukturen vars verktygslåda invigdes den 7 oktober.

SWE-CLARIN är den svenska delen av en europeisk forskningsinfrastruktur som vänder sig till alla som jobbar digitalt med språk eller text, oavsett disciplin. Tanken är dels att göra olika digitala språkresurser, som texter, lexikon, inspelningar av ljud och bild, tillgängliga för forskning, dels att utveckla nya analysverktyg.

– Vi har nu nått halvtid av Vetenskapsrådets finansiering av infrastrukturen och tyckte att det var dags att presentera ett antal verktyg för forskarna, förklarar Lars Borin. Han är nationell koordinator för SWE-CLARIN men också föreståndare för Språkbanken, som tillsammans med Svensk Nationell Datatjänst representerar GU i projektet. Övriga medlemmar finns vid universiteten i Stockholm, Uppsala, Lund och Linköping, samt vid KTH, Språkrådet och Riksarkivets Digisam.

Korp kallas Språkbankens huvudverktyg för att forska på stora mängder text som ursprungligen utarbetats för språkvetare men som håller på att anpassas också till andra forskares behov. Ett exempel är retorikhistorikern Jon Viklund i Uppsala som använder verktyget *ordbilder* för att undersöka hur attityderna till vältalighet förändrats i det svenska offentliga samtalet under 200 år.

– Genom att ta reda på vilka grannar ett ord har i en text, exempelvis ordet ”vältalighet”, kan man undersöka hur människor i olika tider såg på själva företeelsen. Ord som hänger samman med retorik befann sig till exempel i ett betydligt mer positivt sammanhang för 150 år sedan än idag.

ETT ANNAT VERKTYG kallas *entitetsuppmärkning* och innebär automatiska metoder att identifiera namn på platser eller personer i olika material.

– Bland annat samarbetar vi med Svenskt kvinnobiografiskt lexikon, berättar Lars Borin. Vi håller på att ta fram en karta där de ingående kvinnornas födelse- och dödsort är markerade men också olika platser där de verkat.

Svenska dialektkartor på sekunden är en tjänst som utvecklats i samarbete med Stockholms universitet och bygger på material från bloggar.

– Den som är intresserad av ett dialektalt ord, exempelvis ”ostkrok”, kan skriva in det i en ruta och genast kommer det upp en karta över Sverige och Finland som visar var ordet används, nämligen främst i Norrland. Vad det betyder? ”Ostbåge” förstås.

> »Men man tänker kanske inte på att även vår historia är skriven på flera språk.«
>
> LARS BORIN

SWE-CLARIN intresserar sig dock inte bara för svenska språket. Med hjälp av verktyget *universal dependencies* kan forskaren göra syntaktiska analyser oberoende av språk.

– ATT SVERIGE ÄR flerspråkigt är väl de flesta medvetna om. Men man tänker kanske inte på att även vår historia är skriven på flera språk. Den som vill läsa Axel Oxenstiernas brev måste exempelvis kunna både sextonhundratalets svenska, tyska, franska och latin. Men språkanalys kan förstås också handla om lexikon. Med hjälp av *länkning av parallella texter* kan man exempelvis studera hur ordet ”stol” översätts i olika texter för att få fram bättre översättningar.

Ytterligare ett hjälpmedel, *textprofilering*, vänder sig till forskare som behöver bedöma svårighetsgrad på olika texter, antingen för att underlätta språkinlärning eller för att stötta barn eller personer med kognitiv problematik som behöver enklare texter.

ATT SWE-CLARIN:S verktyg har utvecklats i samarbete med forskare är viktigt, påpekar Lars Borin.

– Den som exempelvis söker på Google ser ju inte vilken information hen inte får av allt det som finns där. Och det är just det vi kommer att arbeta vidare med, att försöka ge forskarna just det de faktiskt är ute efter så att de inte riskerar att missa viktig information.

TEXT: EVA LUNDGREN
FOTO: JOHAN WINGBORG

SWE-CLARIN

Det stora europeiska infrastrukturprojektet CLARIN (Common Language Resources and Technology Infrastructure) syftar till att göra digitala språkresurser samt språkteknologiska verktyg tillgängliga för forskare inom alla discipliner, men särskilt inom humaniora och samhällsvetenskap. SWE-CLARIN kallas den svenska delen. Vetenskapsrådet har utsett Göteborgs universitet värdinstitution och ger ett stöd på 50 miljoner kronor fram till 2018. Mer information finns på: www.sweclarin.se.

Vill du kolla upp var ett dialektalt ord används? Gå in på: www.ling.su.se/kartverktyg.