

Grammatik program

Språkteknologi

– teorin som gör det möjligt att lära datorn förstå språk

Aarne Ranta är professor i datavetenskap som talar sex språk och därtill har studerat tio. Men när han utvecklar språkverktyg använder han idéerna från maskinernas språk för att bygga grammatiker för datorn.

– Det gäller att se överlappningarna, säger han. Det ena inspirerar till det andra.

Samarbetet mellan institutionerna Data- och informationsteknik och Lingvistik etablerades redan på 90-talet med professorerna Robin Cooper och Bengt Nordström. För tre år sedan växte samarbetet ytterligare med Institutionen för Svenska språket och professor Lars Borin. Tillsammans bildade trion Centrum för språkteknologi som i dag är väl etablerat. Grammatiker som mjukvarubibliotek blev centrumets första gemensamma projekt med Vetenskapsrådet som finansierar, där de utvecklade IT-verktyg för att skapa språkresurser. Språkbanken, som är mest känd för Svenska Akademiens ordlista, var föremål för insatserna i projektet och ur den teknik som uppstod i projektet kunde forskarna på Data- och informationsteknik senare utveckla tekniken bakom Saldo.

Öppet lexikon

Saldo är ett stort open-source-lexikon som öppnade i maj 2008. Det är en uppdaterad och utvidgad version av Svenskt associationslexikon, SAL, med rötter i 70-talet. Aarne Ranta gillar att Saldo är ett öppet och levande verktyg som används och är till nytta för många språkanvändare varje dag. Namnet Saldo kommer från förkortningen SAL med tillägget do som betyder två på hindi.

Mogen forskningsmiljö

Tekniken inuti Saldo heter funktionell morfologi och har stark anknytning till funktionell programmering, som länge varit ett styrkeområde inom Institutionen för data- och informationsteknik. Morfologi betyder ordformer eller formlära och är den del av lingvistik som studerar ordens struktur, deras former och bildning. De båda programteknikerna menar Aarne är ett resultat av forskningsmiljön på institutionen som han beskriver som mogen och utvecklad.

– Här finns kollegor och masterstudenter från en mängd länder som pratar många språk och hjälper oss. Forskare från hela världen kommer hit och jobbar tillsammans. För doktoranderna blir förutsättningarna att lyckas så mycket större när de är integrerade i miljön och har fler forskare och andra doktorander att knyta an till. Mångfald är bra för institutionsmiljön, men det tar lång tid att utveckla en sådan miljö, säger Aarne Ranta.

Lär datorn grammatik

Språkteknologigruppen har haft programspråket Grammatical Framework som genomgående tema det senaste decenniet. Grammatical Framework är ett programspråk designat för grammatik som översätter till flera språk samtidigt, och där orden uttrycker samma innehåll på de olika språken. För att ge en bild kan vi jämföra med ett översättningsprogram som många mött, Googles översättning av webbsidor. Googles översättning har en stor fördel: den kan översätta vilken text som helst. Men Googles verktyg ”kan” inte grammatik, det är frasbaserat och översätter ord för ord. Eftersom översättaren också bygger på statistik, så ”vet” programmet att det ska

mmering!

översätta “a house” till “ett hus” och inte “en hus”. Grammatical Framework kan översätta hur långa ordföljder som helst och ”vet” att det heter ett hus, tack vare den grammatikbaserade tekniken, men har nackdelen att bara kunna översätta det som finns med i grammatiken.

Fråga kartan om vägen

Ur teorierna bakom Grammatical Framework har gruppen senare utvecklat flera språkapplikationer i samarbete med dialoglabbet på Campus Lindholmen. Digitala kartor som man frågar om vägen och där kartan sedan visar hur man ska åka, är ett exempel på taligenkänning som gruppen gjort.

För språkteknologin är det förstås utmanande att man jobbar med något så levande och föränderligt som språket.

– För att komma runt detta har vi områden som vi utvecklar och där vi kan översätta med stor precision, som spårvagnar och hållplatser. Matteövningar har varit ett annat bra område att utveckla tekniken med.

Från 12 till 23 språk

I sommar arrangerar institutionen Resource Grammar Summer School. Aarne hoppas och önskar sig att det kommer två deltagare från varje europeiskt språk så att Grammatical Framework kan växa och utökas till att kunna översätta alla 23 officiella EU-språk. Än så länge hanterar programmet de stora europeiska språken plus bulgariska, danska, finska, norska och svenska.

Apropå EU så kommer vi in på finansiering. Det är arbetsamt att söka EU-medel, tycker Aarne. Men han anser att det är en finansieringsform som lönar sig väl, eftersom man får samarbete på köpet. Och då är vi tillbaka till samtalets början. Samarbete och mångfald är nyckelord för språkteknologin vid Göteborgs universitet. Utan dessa hade området inte varit den vetenskapliga framgång som det är idag.

Markus Forsbergs avhandling “*Three Tools for Language Processing: BNF Converter, Functional Morphology, and Extract*” presenterar tekniken i Saldo.



Aarne Ranta, professor i datavetenskap, med språkteknologi som specialitet.

Text: Åsa Ekvall