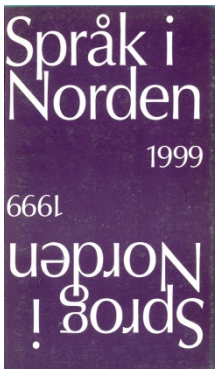
Sprog i Norden

Titel: Informationsåtkomst på flera språk

Forfatter: Jussi Karlgren

Kilde: Sprog i Norden, 1999, s. 37-43

URL: http://ojs.statsbiblioteket.dk/index.php/sin/issue/archive



Søgbarhed
Artiklerne i de ældre numre af Sprog i Norden (1970-2004) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Informationsåtkomst på flera språk

Jussi Karlgren

Att hitta information kan vara knivigt. Det kan vara så att den som söker information vet exakt vad den vill ha fram, men inte har precis klart för sig var det finns; det kan också vara så att den som söker inte riktigt vet vad som finns men har en känsla av att någon sorts hjälp finns att få, bara frågan är rätt ställd. De senaste millennierna har människor lagrat information på externa lagringsmedia av olika slag: det finns mer och mer information att tillgå, men av skiftande kvalitet, otydliga ägarförhållanden, oklar provenans och det är mindre och mindre tydligt vem läsaren kan fråga till råds för att hitta rätt.

Det finns en mängd olika tekniker för att hjälpa folk hitta information. Hyllor och ordentligt markerade bokryggar är ett gott första steg, alfabetisk eller någon annan systematisk hyllordning ett ytterligare, kortkataloger för tillexempel ämnesordsregister med handskrivna nyckelord som ger andra sorteringskriterier än hyllorna ett tredje. Ju fler olika sorters index, desto lättare att hitta grejerna, och desto arbetsammare att adminstrera och upprätthålla. Det är naturligtvis här datorer kommer in. Biblioteken arbetar i dag med tekniska hjälpmedel för kataloghantering, och informationsteknologin används just för det den är bäst på: att adminstrera stora mängder information och sprida den med väldigt låg marginalkostnad – allt vilket oftast anses vara bra.

Digitala bibliotek

Ett bibliotek inbegriper mycket mer än bara tekniker för att hålla information tillgänglig. Urval, införskaffning, arkivering, datamagasinering, standarder för arkivbeständiga format och rättsliga frågor är bara några av de tekniska frågorna; frågor om bibliotekens roll i ett scenario där inte enstaka exemplar av

textdokument är unika, utan informationen är duplicerbar och åtkomlig överallt är lika så viktiga.

Biblioteken har nu inte bara texter. Information finns lagrad på flera olika sätt. Mycket finns i text, men också bilder, både stillbilder och rörliga bilder; ljud av olika slag, både tal och annat; material på olika språk, där inte bibliotekspersonalen förstår alla språken. Det här heterogena materialet används av en mängd olika anledningar: folk vill arbeta med en avhandling, roa sig en stund, skriva en artikel, lösa ett korsord, släktforska – och alla behöver olika sorters svar på sina sökfrågor. Allt det här vet biblioteken sedan länge, och har mer eller mindre välfungerande rutiner för att hantera diversiteten. Utan motsvarande rutiner och insikter är inte en digital samling informationsbärande dokument något digitalt bibliotek.

> Det är mer än bara en digital kortkatalog som behövs för att ett bibliotek är digitalt; det behövs mer än databas av texter för att bygga ett digitalt bibliotek.

Informationsåtkomst med hjälp av datorer

Informationsåtkomst med datorer är till sin enklaste form ett kortregister inmatat på en dator. Så fungerar de flesta bibliotekssystem idag. I mer avancerad form skapas registret automatiskt genom automatisk granskning och tabulering av texterna. Så fungerar de sökmotorer som finns på Internet.

Informationssystemets bild av texten och dess innehåll är ytterst enkel: en påse ord, i princip. Systemet granskar texterna, räknar orden, väger ovanliga ord mot vanliga, viktar långa dokument mot korta och ordnar ett index över texterna, viktade efter ordförekomst. Den som söker text förväntas uttrycka sig med en sökfråga bestående av några enkla ord som paras ihop med orden i indexet, och de dokument som har flest lika ord med sökfrågan erbjuds användaren som tänkbart relevanta. Inte särskilt mycket till språklig sofistikation, med andra ord. På ett visst plan är teknikerna i stort sett språkoberoende.

Men det finns flera ganska långtgående antaganden om språk som ligger till grund för den här sortens systemsdesign. Främst

och viktigast att ordförekomster är ett gott mått på texters innehåll; därnäst att användare kan lista ut vilka ord som kan användas, och ytterligare att ord i användarens sökfråga är ett gott mått på användarens sökbehov. Alla antaganden är problematiska.

Antagandena bygger på att språklig vaghet är ett brus som skall reduceras i möjligaste mån i analysskedet, så att sökprocessen blir så rätlinjig som möjligt: om ordens valör och position i språket räknas ut i förväg blir själva sökningen inte en språklig fråga om betydelse utan en datalogisk fråga om vektorjämförelse. Det är en rimlig förmodan att flertydighet kan reduceras genom studier i språket i förväg, och att enkla och rensade modeller sedan kan användas vid söktillfället – men då får systemet helt bortse från att texters alla tänkbara användningsområden och ordens alla betydelsefasetter inte alls går att förutsäga. En samling kåserier om Blekingegatan i Stockholm kan vara högst relevant för den som vill läsa mer om Greta Garbo utan att hennes namn någonsin nämns; en artikel om ångfartyg skriven på artonhundranittitalet kan vara relevant för den som vill planera kollektivtrafik på nittonhundranittitalet. Ord är ganska klena enheter för att etikettera texter med.

Och för vi in en tidsaxel i resonemanget är det ju så att ord dyker upp i språket hela tiden, och försvinner i nästan samma takt. Ords valör och betydelse förändras från år till år och från vecka till vecka. En ny artikel om ångfartyg är förmodligen mer inriktad på nostalgi och restaurangmenyer än en gammal artikel som kanske behandlar transportbehov; en femtio år gammal artikel om elektronrör kan komplettera en tio år gammal artikel om transistorer alldeles adekvat trots att ordförekomsterna knappt överlappar.

Vi har bara börjat förstå relationen mellan ord och betydelse i text. Det krävs mer forskning om ord i bruk.

Jussi Karlgren

Språkoberoende system? – Att söka i nordiska texter

Under det skenbart icke-språkliga ligger det ju ganska mycket språklig kunskap inbäddad i den statistiska språkmodellen: ord är inte alltid bara ord. En del språk böjer sina ord så de varierar i utseende från gång till gång. Engelska gör det mycket knapphändigt, svenska en hel del mer, och del språk i Norden gör det riktigt gärna och mycket. En del språk använder sammansatta ord: engelska gör det ytterst ogärna, och orden tenderar vara invarianta även när de ingår som delar av större betydelseelement; svenska gör det mycket gärna, liksom andra språk i Norden. Det gör att enkla ordsökningar på svenska kan snava över böjningsformer eller sammansättningar. Att skriva "veckomatsedel för skolor i Stockholm" i ett sökfönster ger en inte texten "Matsedel för Katarina norra skola vecka 48" om inte systemet vet någonting om formlära.

Engelska är olikt språk i Norden. Inte väsensskilt: det fungerar att leta efter svensk text med system som är utvecklade för engelska, men sämre. De flesta verkar ta det för givet, som om icke-engelsk text av naturen är bökigare och svårare att handskas med: det är precis som med prickar och ringar och cediljer – internationellt bråte som bara stökar till tangentbord och skrivare. Men det är inte givet alls att det skall behöva vara så. System som grundar sig på antaganden om språkliga teckens invarians och relativt fast ordföljd kommer att ge sämre resultat för material som är på språk som hänger sig åt sammansättningar och andra produktiva avledningsmekanismer eller vidlyftig böjningsmorfologi och därmed sammanhängande uppluckrad ordföljd. Men de antagandena behöver inte vi göra i Norden.

Och faktum är att engelskan tillåter systemen ta en förledande enkel genväg till analys av texters innehåll genom att en tabell med enkla grafiska ordförekomster ger en bättre innehållsanalys på engelska än på många andra språk. Oförtjänt bra, om man så vill – för det är studier på engelskspråkigt material som är den empiriska grunden för de ganska starka antagandena om ords betydelse för betydelse som ligger till grund för systemarkitek-

turerna. För att komma längre krävs det studier i texter, textualitet och texters mening – och det faktum att vi stöter på problemet snarare i våra språkområden än i det engelska bör vi ta som uppmuntran. Vi måste arbeta med en mer elegant innehållsanalys för att kunna förstå nordiska texter. Det har vi igen när vi sedan åter tittar på engelska!

Språk är i allmänhet olika, och speciellt är engelska olikt många andra språk. Det är till fördel, om vi vill arbeta med språkteknologi i Norden.

Att hitta material på flera språk

Information finns på flera språk, men vi bor i små språkområden med relativt litet informationsutbud. Det är en rimlig strategi även när vi inte primärt söker eller letar efter någonting på något specifikt språk att första hand vända oss efter engelskspråkigt material.

Det här är naturligtvis en självuppfyllande förutsägelse: vill vi folk ska hitta våra verk ser vi till att de blir skrivna på engelska, och det gäller även våra grannar och närmaste kollegor. Den lokala publiken är lika internationell som mer fjärran gäster. Och i förlängningen osynliggör detta nordiskspråkig information.

Det behöver nu inte gå så illa. Om vi väljer att betrakta sökning i de flesta fall som språkoberoende kan vi se till att bredda den automatiskt. Om vårt system känner till att vi kan läsa svenska, danska och flera sorters norska utan större problem och med visst besvär även tyska, är det inom möjligheternas ram att se till att sökfrågan hämtar dokument på flera språk. Det finns flera tekniska lösningar på hur, men alla grundar sig i insikten att det inte är särskilt komplicerat att översätta enstaka termer i en sökfråga eller i ett index. Inte komplicerat, men resurskrävande: det krävs omfattande flerspråkiga termsamlingar, textsamlingar och textstudier. Och sen blir det fel ganska ofta med dagens tekniker, men det blir det ändå i söksammanhang.

Det går att göra nordiska texter synliga i ett flerspråkigt sammelsurium. Det kräver mer terminologiskt arbete.

Jussi Karlgren

Enstaka dokument är inte alltid svaret, och dokument har mer än bara innehåll

Resonemanget hittills har handlat om att söka efter texter och har utgått ifrån att ett duktigt system som svar på en sökfråga levererar en mängd texter varav de flesta är relevanta i någon teoretisk mening. Det är inte särskilt hjälpsamt, egentligen. Ibland kan ett kort svar vara tillräckligt, ibland är det användbart med en sammanfattning av flera texter, och ibland är det något litet faktum i en väldigt stor mängd texter som skall excerperas och sammanställas i en rapport. För den här sortens arbete krävs det mer studium av texters egenskaper generellt, och texters egenskaper på specifika språk i specifika genrer speciellt.

Och idag har informationssystemen förenklat bilden av dokument så långt att den är ytterligt fattig: dokument är mer än bara påsar av ord – de har stil, tillhör traditioner av olika slag, är producerade med olika avsikter och baktankar, och kommer att tilltala olika läsare vid olika tillfällen. Den sortens distinktioner görs svårare när texterna ligger i uniform i ett informationssystem istället för att stå inbundna, häftade eller stencilerade i en hylla.

Det är tydligare och tydligare att informationshanteringssystem nått en sorts tak för hur väl de kan fungera: forskningen idag inriktar sig mer och mer på att försöka förstå hur användaren tänker och varför; nya system idag tenderar att skräddarsys för specifika uppgifter och användargrupper snarare än att byggas för att hantera alla texter för vem som helst.

Och det är uppenbart att det finns mer att hämta i snittet mellan informationsteknologi och bibliotek. Experimentella system kan redan idag enkelt rekommendera litteratur genom åtkomststatistik: dokument som ofta används tillsammans hör nog ihop på något sätt. Det finns också system som särskiljer olika sorters text – inte på grundval av innehåll utan på grundval av stil. Och det finns system som beter sig olika beroende på förmodad uppgift som användaren sysselsätter sig med för tillfället.

Vi är på väg in i dokumenten. Det kräver mer forskning om språket i bruk och bruket av text.

Vi har bara börjat förstå varför folk läser text, och hur de uttrycker sig när de vill leta efter någonting. Det krävs mer forskning om folks informationsbehov.

Vad behöver forskningen?

Nästa alla punkter som behandlats ovan kommer att kräva mer kunskap om språk om vi arbetar vidare med dem: kunskap om språk i allmänhet och språk i Norden i synnerhet. För att åstadkomma system som hanterar språk i Norden behöver vi ansenliga textsamlingar att experimentera med, vi behöver tillgång till folk som vill använda dem för att studera dem och deras behov, vi behöver studera texter i allmänhet och nordiska texter i synnerhet och vi behöver jämföra system och algoritmer systematiskt med varandra. Och för att utveckla flerspråkig sökning behövs det lexika av olika slag och fler och djupare studier om fackspråk och fackterminologi.

Det finns också en stor mängd allmänna språkvetenskapliga frågor som måste arbetas med för att kunna bidra till språkteknologiska verktyg. Språkvetare måste lyfta blicken från meningar och börja studera text systematiskt; filologer måste lyfta blicken från enstaka dokument och studera generella frågor om texter. Vad är en text? Vad är information i text? Och vad är det för relation mellan text och icke-textuell information?

Det finns ännu större politiska frågor att behandla. Vad vill vi göra med lokala språk? Behöver vi dem? Vem har glädje av dem? Och ekonomiska frågor. Vad är kostnaden för att utbilda stora delar av en hel befolkning i ett annat språk?

Språkteknologin utvecklar verktyg. Vår fråga är vilka verktyg kan vi bygga givet de vetenskapliga, politiska och ekonomiska begränsningar vi arbetar under?