Hur kan informationsystem fås att söka efter information på flera språk?

Jussi Karlgren

30 januari 2001

Idag ska jag prata om hur söksystem och sökmotorer kan fås att arbeta med texter och dokument på flera språk. Det blir fler och fler dokument elektroniskt tillgängliga på fler och fler språk – från att materialet på internet och i informationsteknologiska sammanhang i allmänhet varit uteslutande på engelska går vi mot en textsamling, ett digitalt bibliotek, som bättre avspeglar den flerspråkiga värld vi lever i. Och det icke-engelska materialet växer snabbare i storlek än det engelska. Förutom svenska och engelska som båda redan har sin plats i svensk informationsvardag kommer EU-samarbetet och Östersjö-samarbetet föra in fler språk för oss att läsa på.

Men det betyder inte att allt material är lika synligt, även om det tillgängliggörs på liknande sätt. De flesta sökningar som görs på internet från svenska datorer görs på engelska: läsarna tror de får fler eller bättre träffar på det sättet. Det är förstås en självuppfyllande förutsägelse. Så där borde vi hjälpa till, genom att bygga maskineri som klarar flera språk.

De system som finns är oftast enspråkiga: förment språkoberoende teknik men konstruerade från början och fortsatt trimmade för att hantera engelska språkliga egenheter – särskrivna sammansättningar, futtig formlära och fix ordföljd. De är inte ägnade att hantera ryska, latin och svenska med samma kompetens. Det går dock att ordna med bättre algoritmer för texttröskning, och det är vi flera som jobbar på.

Det är svårt att fråga på ett främmande språk. Men det behöver inte vara ett problem. Att få ett system att översätta en sökfråga till flera språk är inte alls tekniskt omöjligt. System som relaterar "fiskmås" till "seagull" och "larus" (likväl som till svenska näraliggande termer som "trut" och "tärna"!) finns i laboratorierna.

Viktigare är att själva gränssnittet inte stödjer flerspråkighet. Systemen numera brukar tillåta användaren välja *ett* språk. De flesta svenskar läser gärna svenska, utan större besvär engelska, stavar sig om det är nödvändigt igenom danska och norska, och om det är riktigt viktigt kan de dra sig till minnes nåt ur skoltyskan eller skolfranskan. Hur ska ett gränssnitt som förklarar det där se ut? Hur ska läsaren tala om för systemet vilka språk som är acceptabla? Det är inte svårt, men måste ordnas.

Men det knivigaste är nog att hjälpa läsaren avgöra om texten är bra om

den är skriven på ett främmande språk. Det första man som läsare på ett främmande språk förlorar är möjligheten att se om en text är trovärdig och tillförlitlig. Det behövs bedömningshjälp: system som hjälper en fatta beslut om texten är läsvärd eller inte. På vilka grunder? Det är en forskningsfråga. Men vi vet väldigt lite om hur folk fattar läsbeslut själv, ens enspråkigt.

Så söksystem för flera språk måste

- understödja interaktion med texter på flera språk,
- hjälpa och uppmuntra formulera sökfrågor på flera språk med avstamp i ett,
- kunna söka i andra språk än engelska, och
- hjälpa folk läsa och bedöma texter på andra språk.

Allt det där är nödvändigt – och kommer inte bara att ge oss flerspråkiga söksystem utan även bättre enspråkiga sådana. Och om inte vi i små språkområden tar initiativet i att utveckla system av det här slaget kommer resultatet helt säkert inte vara till vår fördel!