TORE AHLBÄCK

Indexspråk och tesaurusproblematik

Ahlbäck, Tore, Indexspråk och tesaurusproblematik [Index language and problems of thesauri] Kirjastotiede ja informatiikka 4(2): 53–57, 1985.

The purpose of the article is to discuss the problems involved when using and constructing thesauri. The main conclusion is that the future of conventional indexing methods is obscure, and that the best alternative is to start with free language indexing and searching – as soon as possible.

Address: The Donner institut, PB 70, SF-20501 Åbo, Finland.

I denna artikel är det min avsikt att diskutera ett begränsat men svårtacklat problem i samband med konstruktion av tesaurer, nämligen begreppet *'related terms'*, och dra vissa slutsatser av detta.

Enligt *Guidelines for the establishment and development of multilingual thesauri.* 2. rev. ed. Prep. by Derek Austin & Peter Dale. Paris: UNISIST, 1981. (PGI/81/WS/15) är en *thesaurus* »the vocabulary of controlled *indexing language* /– – –/, formally organised so that the *a priori* relationships between concepts (e.g. as 'broader' and 'narrower') are made explicit.» (s. 7) Definitionen på indexspråk är följande: »A controlled set of terms selected from natural language and used to represent, in summary form, the subjects of documents.» (Ib.) Den här utnyttjade standardens uppfattning om *a priori*-relationer skall ytterligare redovisas: »Those *a priori* or thesaural relationships between terms assigned to documents and other terms which, because they form part of common and shared frames of reference, are present by implication. /– – –/ 'Banks' would imply a broader term such as 'Financial institutions'; 'Computers' is mentally associated with 'Data processing'; and 'Amsterdam' implies the wider location 'Netherlands'. Any of these mentally-associated terms might serve as a user's approach to the subject index. These relationships are document-independent, since they are generally recognised and could be established through reference to standard works, such as dictionaries and encyclopaedias.» (s. 2) Den bild som ges av sambandet mellan de här anförda termerna är dock långt mera problematisk än det anförda citatet ger vid handen.

Inget problem vidlåder de hierarkiska relationerna, de är baserade på solid och välkänd klasslogik, dvs de utgör inget problem på det principiella planet. Den enda invändningen som kan anföras, är mot benämningen »mentally-associated» som en beskrivning av sambandet mellan termer som står i ett hierarkiskt förhållande till varandra. Den använda benämningen inducerar föreställningen om en liknande godtycklighet ifråga om »broader terms» och »narrower terms» som de facto förekommer ifråga om »related terms». Så är icke fallet. Svårigheten vad gäller de hierarkiska termerna består framförallt i att få klart för sig hur i det enskilda fallet klasser respektive begrepp förhåller sig till varandra. Svårigheten kan uttryckas på följande sätt. För att exakt fastställa förhållandet mellan klasser och mellan begrepp så måste klasser och begrepp språkligt vara *entydigt* bestämda. Uttrycks

betydelse bestämms med hjälp av definitioner. Dessa kan i sin tur vara för vida eller för trånga eller både för vida och för trånga – i de fall då de inte är adekvata. Ett annat problem i samband med de hierarkiska relationerna består i att finna en enhetlig indelningsgrund för indelning av en klass i underklasser. Ett klassiskt exempel på svårigheterna härvidlag illustreras av följande gammalkinesiska indelning av djur. Dessa indelas i bl.a. följande grupper: djur som tillhör kejsaren, balsamerade djur, tama djur, spädgrisar, fabeldjur, vilda hundar, djur som just slagit sönder en kruka, djur som på avstånd ser ut som flugor.

De hierarkiska relationerna utgör inte ett problem på det teoretiska planet. Det gör däremot de relaterade termerna. Detta antyds i UNISIST-standarden med följande ord: »The associative relationship. /– – –/ This is the most difficult of the basic relationships to define in terms of positive rather than negative characteristics. It covers relationships between pairs of terms which are not members of an equivalence set, nor can they be organized as a hierarchy in which one term is subordinated to another, yet they are *mentally associated* to such an extent that the link between them should be made explicit in the thesaurus, on the grounds that it would reveal alternative terms which might be used for indexing or retrieval. This relationship is reciprocal, and is indicated by the abbreviation '*RT*' (related term), or its equivalent in other languages.» (s. 39. Kurs T.A.)

Vad innebär begreppet 'mentally associated'? I varje fall innebär det någonting som inte kan redovisas i explicita termer och förklaringen är enkel. Relationen upprättas inte med hjälp av logiska regler som fallet är ifråga om de hierarkiska relationerna utan de upprättas på mer eller mindre intuitiv väg, m.a.o., här ingår ett subjektivt moment. Detta är naturligtvis inte någon hemlighet för standardens upphovsmän, som försynt ger följande uppmaning till den som avser att konstruera relaterade termer i en tesaurus: »It is important to exercise strict control over the choice of terms linked in this way, and to avoid *subjective* judgements.» (Ib. Kurs. T.A.) Rådet är välmenande men troligen omöjligt att följa.

Standarden inducerar dock uppfattningen att det trots allt skall gå att upprätta denna typ av relationer med hjälp av regler. Vi skall kasta en blick på arten av de regler som standarden redovisar för. Som allmän riktlinje fastställs: »/– – –/ that one of the terms should be strongly implied, according to the frames of reference shared by the users of an index, whenever the other is employed as an indexing term. More specifically, it will frequently be found that one of the terms is a necessary component in any explanation or definition of the other, to the extent that the term 'Birds', for example, forms a necessary part of the explanation of 'Ornitology'.» (Ib.) Nu kunde man vänta sig att det anförda exemplet »ornitology – birds» skulle klargöra vad som menas med att den term som väljs som relaterad term, i detta fall »birds», skall ingå i tesaurus-användarens referensram. Det anförda exemplet demonstrerar däremot en typisk hierarkisk relation, dvs »birds» är en nt-term till »ornitology». Emedan standarden framhåller denna relation som ett typexempel på en rt-relation uppstår hos läsaren en viss förvirring.

Standarden hävdar att termer som kan stå i ett rt-förhållande till varandra antingen båda kan tillhöra samma »kategori» eller tillhöra olika »kategorier». Eftersom någon tredje möjlighet inte gives (lagen om det uteslutna tredje), följer härav att det egentligen inte finns några gränser för området inom vilket man kan välja rt-termer. Standarden lämnar dock inte läsaren i sticket, utan ger i verkligheten exaktare uppgifter om hur de relaterade termerna skall upprättas. Först redovisas för hur rt-termer tillhörande samma »kategori» som utgångstermen skall väljas. Den första av dessa regler upplyser om att det går att utnyttja en term, med en betydelse som delvis täcker utgångstermens betydelse, som rt-term. Det konkreta exemplet är »Boats» som rt-term till »Ships». Här kan påminnas om att hierarkiska relationer konstrueras meddelst antingen division, dvs indelning av en klass i underklasser, eller partition, dvs indelning av ett helt i delar. Om det är fallet att »boats» och »ships» tillhör samma klass (innebörden i standardens term »category» är inte helt klar) kan deras inbördes relation vara följande: det ena begreppet är antingen överordnat, underordnat eller sidoordnat det andra. Om det inte går att fastställa den inbördes relationen, kan antagandet om gemensam klasstillhörighet sättas ifråga. Ponera att »ships» och »boats» är sidoordnade. I så fall vore det korrektare att benämna deras inbördes relation på detta sätt än att tala om en »mentally associated» relation. Vinsten skulle vara att

alla genast skulle inte vad det är fråga om. (Det kan här inskjutas att »boats» i svenskt språkbruk är underordnat »ships».)

Problemet med termer som delvis täcker varandra, dvs kvasisynonymer, borde naturligtvis i första hand lösas genom att endast den ena av två dylika termer används i indexspråket, medan den andra får en use-hänvisning.

Detsamma som här sagts om den första regeln gäller även för den andra: »Concepts linked by a familial or derivational relationship (i.e. one of the concepts was derived from the other), can also be regarded as belonging to this group. This would apply to terms such as 'Hinnies' and 'Mules', which represent kinds of crossbreed between 'Horses' and 'Donkeys'.» (s. 40) Det anförda exemplet anger med all önskvärd tydlighet att det här är fråga om en hierarkisk relation *par préférance*, nämligen en generisk relation. Varför det här skulle vara motiverat att tala om en »mentally associated» relation låter sig helt enkelt inte helt lätt inses.

Det hade varit att vänta att standarden här hade kunnat utnyttja Wittgensteins ide om familjelikhet, som går ut på följande: AB och BC har B gemensam, BC och CD har C gemensam, och därav följer en »familjelikhet» mellan AB och CD.

Det framgår således, att de regler som anförs som vägledande för valet av rt-termer från samma »kategori» som utgångstermen tillhör, delvis handlar om hur man upprättar hierarkiska relationer, underförstått att sidoordnade relationer räknas till hierarkiska relationer – och det gör de.

Anvisningarna för val av rt-termer ur annan »kategori» än utgångstermen är legio, och därtill inte avsedda att vara uttömmande – vilket är begripligt med tanke på det bokstavligt talat obegränsande området, dvs alla de begrepp som utgångstermen inte har någon »kategori»-gemenskap med.

Regel 1 har följande ordalydelse: »a discipline or field of study and the objects or phenomena studied». Exemplet här är »Forestry» som får rt-relationen »Forests». Detta råd verkar förnuftigt, men dess värde begränsas av följande omständighet. Inom ramen för indexering av vetenskaplig litteratur tillhörande en bestämd disciplin kommer huvuddelen av litteraturen att tillhöra denna disciplin, tex sociologi. Att anföra sociologi som rt-term är därmed av lätt insedda skäl oändamålsenligt. Standardens regel har därmed relevans närmast för indexering av stora övergripande ämnesområden, däremot inte för indexering av bestämda avgränsade discipliner.

Det ges ytterligare åtta anvisningar eller rättare sagt tips för hur rt-relationer mellan begrepp tillhörande olika kategorier» skall upprättas. Anvisningarna är följande: »An operation or process and its agent or instrument», »an action and the product of the action», »an action and its patient», »concepts related to their origins», »concepts linked by causal dependence», »a thing and its counter agent» samt »syncategorematic phrases and their embedded nouns». (s. 42s) Utan att gå in på detaljer kan som allmänt omdöme sägas att det inte framgår varför just dessa anvisningar har valts; anses dessa vara uttömmande och tillräckliga och i så fall på vilka grunder anses detta. Mitt intryck är snarare att det är fråga om råd som i verkligheten kunde mångfaldigas utan att man ändå skulle kunna uttömma möjligheterna att bilda rt-termer mellan begrepp tillhörande olika »kategorier». Om de anvisningar som ges på denna punkt i standarden kan därtill lätt konstateras, att de i första hand framstår som tillämpliga inom naturvetenskap och teknik. Den allvarligaste bristen består dock i att de givna anvisningarna är så trubbiga. Ta exempelvis anvisningen om att upprätta en rt-relation i enlighet med »a thing and its counter agent». Det exempel som ges för att illustrera detta lyder: »PLANTS *RT:* HERBICIDES» (s. 43) Jag tror mig förstå *varför* denna anvisning ges, nämligen på grund av att artiklar växtbekämpningsmedel rimligtvis också implicit handlar om växter. Men i verkligheten överlämnas genom detta sätt att indexera en stor, en alltför stor del av arbetet med att klargöra relevansfrågan åt användaren att utföra. Det hade varit indexerarens sak att indexera en artikel om växtbekämpningsmedel *även* på örter om artikelns innehåll verkligen motiverade detta.

Rt-relationer uppgjorda på det sätt som standarden föreslår kan utan tvekan ha ett värde för indexeraren, men om de också har det för användaren, så tyder det på brister i indexeringen. För användaren borde enligt mitt förmenande rt-termer anges endast i form av sidoordnade termer till utgångstermen. Då skulle det åtminstone gå att ge en teoretisk motivering till arrangemanget med bt-, nt, och rt-termer åt en användare, som får en tesaurus i handen med upplysningen, att det mest sofistikerade sättet att

återvinna information är att utnyttja ett dylikt hjälpmedel. Nu är läget det, att om en användare vill veta enligt vilka principer rt-termerna i tesaurens mikrohierarkier är uppgjorda, så blir svaret, att principerna nog inte är helt klara.

Men en tesaurus som opererar med under-, över- och sido-ordnade relationer kanske inte mera bör kallas en tesaurus, utan snarare ett multiklassifikatoriskt system. Fördelen med ett sådant system, och multiklassifikation av dokument baserad på ett sådant system, är att man åtminstone kommer ifrån den egendomliga luddighet i själva reglerna som styr indexering med hjälp av ett kontrollerat indexspråk, och som i första hand gäller rt-termerna.

Jag är för egen del visserligen inte helt övertygad om vinsten av att ersätta indexering med multiklassfikation av det enkla skälet att jag inte betraktar indexering med hjälp av ett kontrollerat indexspråk som en verksamhet som har framtiden för sig. Indexering av denna typ förutsätter en hög nivå både på indexeringsinstrumentet och på indexerarens kunskaper om det område som indexeringen gäller, två förutsättningar av vilka inte ens den ena alltid föreligger, för att inte tala om båda. Det är en dyrbar sysselsättning vars förekomst kan försvaras enbart om resultatet är tillfredsställande. I fråga om det område, på vilket jag själv kunnat följa med resultatet av indexering i internationellt sammanhang och på vilket jag även själv utfört och utför indexering, nämligen religionsvetenskap och teologi, kan jag kort konstatera att resultatet – och härvid syftar jag inte enbart på mitt eget – *inte* är speciellt imponerande. Dess allra största brist är utan vidare den subjektivitet och godtycklighet som – troligen med nödvändighet – *måste* kännetecknа indexering utförd meddelst ett kontrollerat indexspråk, varvid förekomsten av en tesaurus endast innebär en förbättring i gradavseende, inte i artavseende.

Det är min övertygelse att indexering på basen av det naturliga språket är den enda form av indexering som i framtiden kommer att komma i fråga. För detta talar personella, ekonomiska och tekniska resurser ett entydigt språk. Om det nu är på detta sätt – jag formulerar gärna detta antagande i hypotesens form – vore det då inte motiverat att satsa tillgänglig forskningspotential för att utveckla indexering med hjälp av naturligt språk, och därigenom snabbare uppnå den tid då det enkla permuterade indexet mera allmänt tillhör historien och ADB-baserad maskinindexering nått en högre grad av sofistikering.

Mitt synsätt är inte bara dikterat av att jag anser indexering med hjälp av ett kontrollerat indexspråk inte når upp till en tillfredsställande exakthet utan även att detta sätt att handskas med informationsåtervinning håller på att bli inaktuellt på grund av den allmänna utvecklingen inom området lagring av information. Vi befinner oss i ett skede där lagringen håller på att övergå från pappersbaserad information till elektroniskt lagrad information som är on-line-tillgänglig. Beakta det scenario som F.W. Lancaster för fram i *Libraries and librarians in an age of electronics*, Arlington, Va, 1982. Lancaster gör här ett antagande om att abstract- och indextjänster inom en tidrymd av 20–30 år helt kommer att försvinna i sin konventionella form. Han exemplifierar sitt antagande på följande sätt. Han utgår från en index- och abstracttidskrift, som bevakar en specialdisciplin. Bevakningen görs iform av indexering av artiklarna i 50 tidskrifter med hjälp av ett kontrollerat indexspråk samt uppgörande av abstracts på ett urval av artiklarna. Uppgiften att sätta ihop abstracts förändras dock på grund av allt flere av de 50 tidskrifterna själva börjar förse sina artiklar med abstract, vilka oförändrade intas i index- och abstracttidskriften. Denna förändrar då sin egen verksamhet på det sättet att den börjar lägga upp en databas, i vilken abstracterna och indexorden förekommer i maskinläslig form. Därefter börjar dock alla de 50 tidskrifterna att publicera abstracts till sina artiklar. »The task of 'abstracting', then, involved nothing more than putting the author abstracts into machine-readable form. After the data base had been used for some time, it was recognized that the indexing activity was becoming redundant, since acceptable searches could be done on the text plus abstracts. The human indexing was retained solely to produce a usable printed tool. As the years went by, however, subscriptions to this printed tool declined to the point where on-line income was actually subsidizing the product. The printed tool was then abandoned.» (s. 189) Efter det att den tryckta tidskriften läggs ner fortsätter verksamheten ännu en tid, men när utvecklingen leder till att alla de 50 tidskrifterna blir tillgängliga på on-line, läggs även data-basen ned. De 50 tidskrifterna blir nämligen samtliga till-

gängliga i en egen databas, i vilken det även är möjligt att utföra sökningar.

Även om Lancasters framtidsvision vad tiden beträffar sannolikt inte kommer att uppfyllas vad de nordiska länderna beträffar, är det bara fråga om, just det, en tidsfråga. Det är för vårt resonemang inte centralt om förändringen inträffar om 20 år eller om 50 år, det viktiga är trenden. Mot denna bakgrund verkar det motiverat att välja en form av indexering som är anpassad till den framtidsutveckling som väntar oss – en indexering som utförs elektroniskt på basen av de uppgifter som lämnas i dokumentet av författaren själv – i form av titel och bifogat abstract – och inte en form av indexering där en annan person än författaren analyserar dokumentet och förser det med indextermer. Att fortsätta att ägna sig åt den senare formen av indexering innebär att personella och ekonomiska resurser utnyttjas för en – snart – museal verksamhet.

Hyväksytty julkaistavaksi 14. 5. 1985

Tämän numeron kirjoittajat:

Ahlbäck, Tore, FT, kirjastonhoitaja, Donnerska institutet
Iivonen, Mirja, YK, vs. lehtori, Tampereen yliopisto
Laaksovirta, Tuula H., YL, vt. apul.prof., Tampereen yliopisto
Leikola, Anto, FT, dos., Helsingin yliopisto
Okko, Marjatta, prof., Tampereen yliopisto
Vickers, Stephen, tutkija, IFLA International Office for UAP, West Yorkshire