LexWare®
Djupindexering

Fler och fler böcker, tidskrifter, artiklar, dokument finns idag i digital form på många bibliotek, förlag, tidningar, och högskolor. Dessa dokumentsamlingar växer snabbt och behöver göras sökbara för sökmotorer. Detta kan göras på olika sätt och med olika ambitionsnivå: från enkel fritextsökning till sökning på innehåll. Sökning på innehåll tillhandahålls t.ex. av Riksdagsbiblioteket, där man kan söka på ett eller flera nyckelord och få fram dokument som handlar just om det utvalda ämnet.

Detta är möjligt tack vore bibliotekarierna som mödosamt läser alla dokument och markerar dem med nyckelord som avspeglar innehållet. Detta kallas manuell indexering. Därefter kan en sökmotor välja de dokument som har de specifika nyckelorden som en användare efterlyser i en sökning. Bibliotekarierna på Riksdagsbiblioteket använder nyckelord från en tesaurus - en ämnesrepresentation med uppdelning i delämne, t.ex. ämnet "fordon" består bl.a. av "personbilar", "lastbilar", osv.

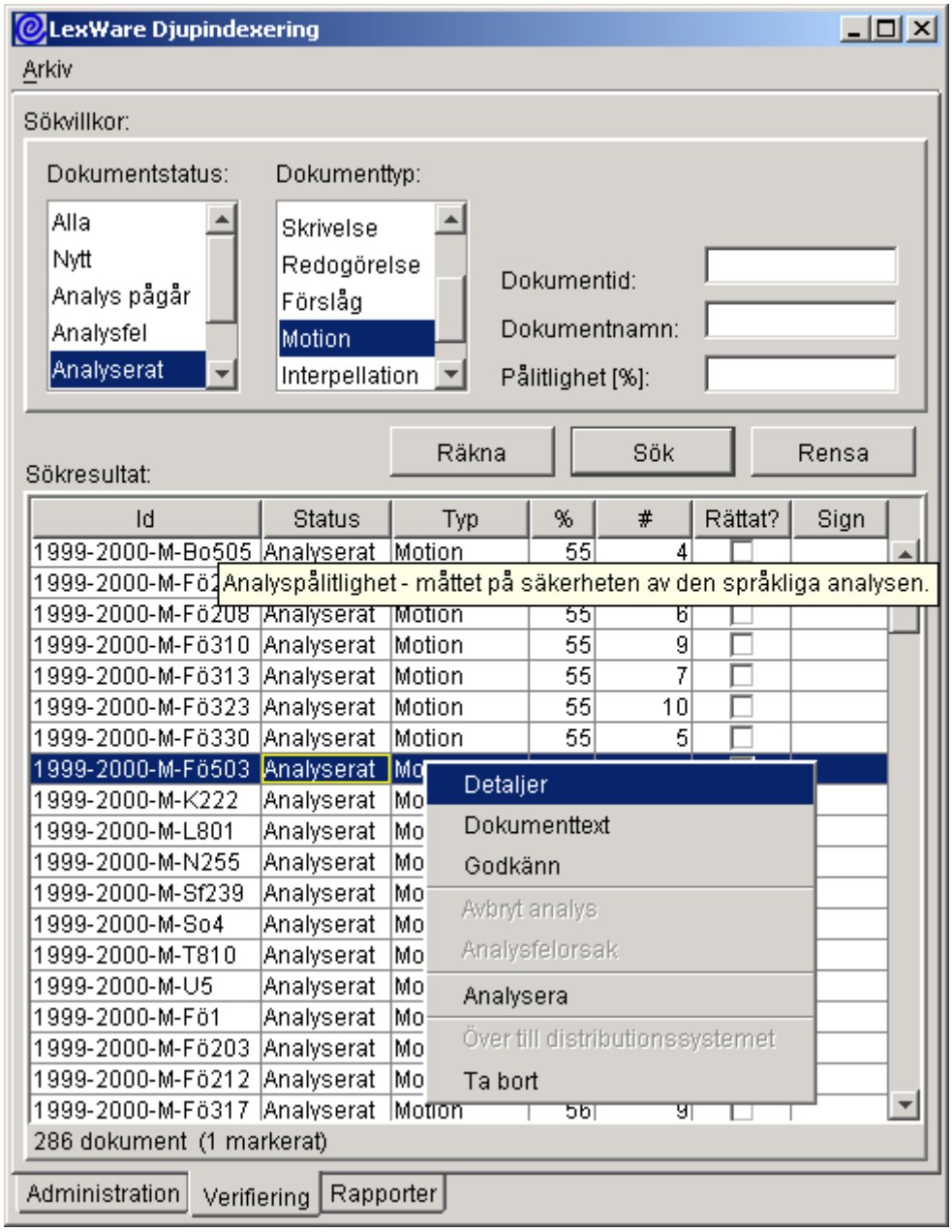
Rätt indexering innebär alltså inte bara att välja huvudämne utan även detaljnivån i ämnesrepresentationen – uppgiften är både mödosam och intellektuellt krävande. Den uppgiften kan utföras nu automatiskt av LexWare Djupindexering. Tack vore systemets kunskaper i svenska och möjligheten till att integrera en extern ämnesrepresentation med LexWare® Djupindexering inbyggda "intelligens" kan systemet analysera svenska texter, och sedan förse dem med lämpliga nyckelord. Tester på Riksdagsdokument har visat att alla automatiskt genererade nyckelord av LexWare Djupindexering är relevanta, och att över 80% av dokument får exakt samma eller mycket lika nyckelord som vid manuell indexering.

LexWare® Djupindexering kan arbeta med eller utan mänsklig övervakning. Vid helautomatiskt arbetssätt skannar programmet en förbestämd indatakatalog för inkommande texter, analyserar dessa och skriver ut nyckelord till en förbestämd utdatakatalog. Vill man ändra, ta bort eller lägga till nyckelord görs detta snabbt utifrån ett grafiskt gränssnitt, som presenteras nedan.

Inte bara nyckelord utan även systemets kunskaper kan uppdateras och utökas. Övervakning av systemet underlättas av kumulerad statistik som kan tas ut på fritt valda parametrar, över fritt vald dokumentmängd, och med mycket korta svarstider. Systemet bedömmer själv egna resultat och anger bedömningen i procent för varje dokument, som s.k. "pålitlighet". Måttet är avsett för att fastslå ett lämpligt gränsvärde för helautomatisk indexering, och för att sortera fram dokument för eventuell verifering. Nyckelord i tillämpningen som exemplifieras nedan - djupindexering av Riksdagsdokument – kommer från en tesaurus på c:a 4 000 termer. Beroende av längd och typ indexeras varje dokument med 3 till 15 nyckelord.

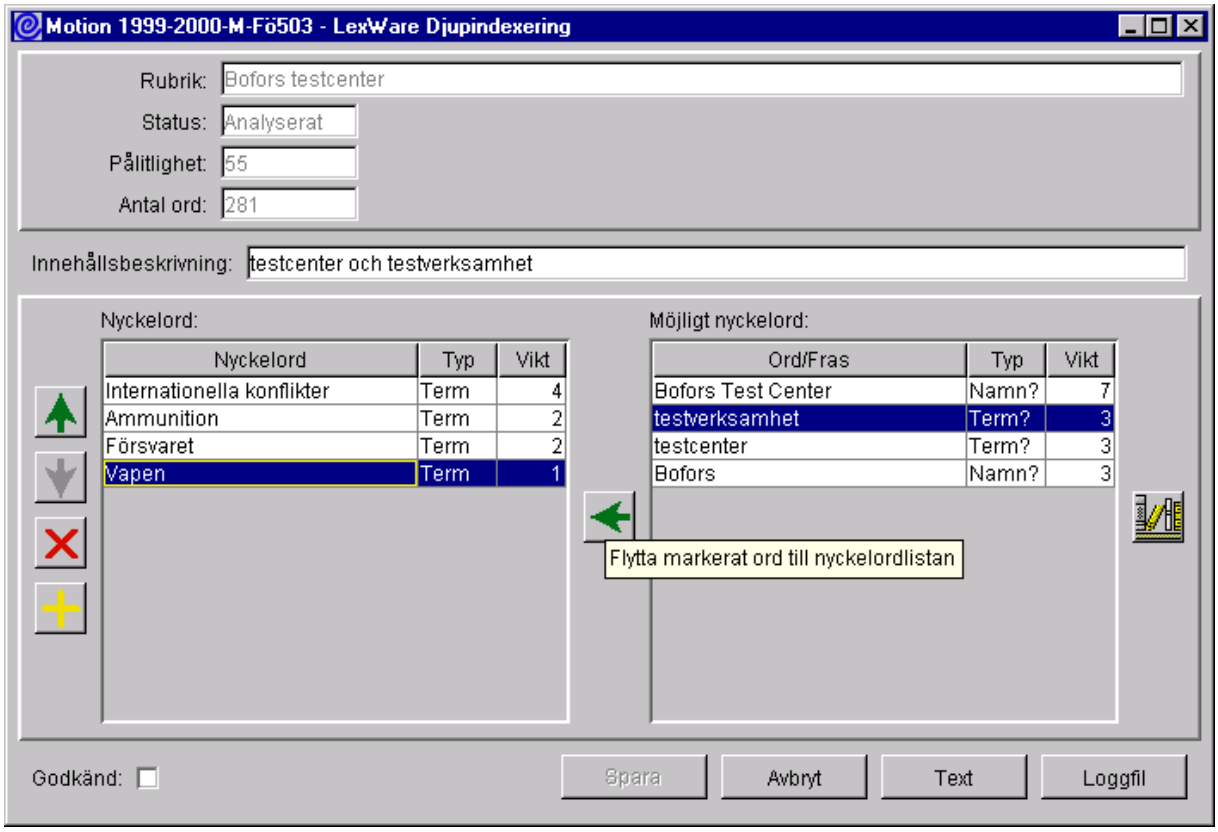
Att överblicka indexeringsresultat

Dokument i LexWare® Djupindexering kan överblickas i olika urval och sorteringar. Nedan visas ett urval av dokument av typen motion som har försetts med nyckelord (status = analyserat). Tabellen med sökresultat kan sorteras på varje kolumn (med en klick/ skiftklick på kolumnnamn). Varje fält är försett med en hjälpbubbla, som kommer upp när markören pekar på fältet; här visas en hjälpbubbla för kolumnen % - pålitlighet. Tabellen innehåller också uppgifter om hur många nyckelord har genererats (kolumn #), om de genererade nyckelorden har ändrats (kolumn Rättat?) och av vem (kolumn Signatur). Man kan godkänna dokument utan att ens titta på dem - det bara att markera dessa i fönstret och välja Godkänn i menyn, den aktiveras med en högerklick. Dokumenttexten visas i en valfri extern ordbehandlare om Dokumenttext väljs från menyn. För att gå över till granskning och ev. redigering av indexeringsresultat väljs Detaljer från menyn.



Att redigera rubrik, nyckelord och tesaurstermer

I övre delen av fönstret för granskning och redigering av indexeringsresultat är visas grundinformation om dokumentet: identiferare, titel/ rubrik, analysstatus, pålitlighet samt textlängd i ord. I nedre delen visas allt som systemet har genererat för det specifika dokumentet:: innehållsbeskrivning i mittfältet, nyckelord i vänstra spalten, och s.k. möjliga nyckelord i högra spalten.



Innehållsbeskrivning väljs av systemet så att varken nyckelord eller textens egen rubrik upprepas, om möjligt, och därför kan den ses som ytterligare en etikett på textens innehåll utöver nyckelorden, eller som huvudrubrik för de dokument som saknar egen rubrik.

Nyckelorden från tesaurusen finns utsatta i vänstra spalten. Uttryck som är viktiga för textens innehåll men som inte omfattas av tesaurusen väljs fram i en separat lista och visas i högra spalten. Om veriferaren anser att något av de föreslagna möjliga nyckelorden bör finnas med i indexeringen flyttas det till vänstra spalten med hjälp av vänsterpilen.

Om något av de nya nyckelorden anses vara lämpligt även som tesaurusterm kan det sparas i en av s.k. användarvokabulärer. Uttryck från användarvokabulärer används av analysatorn som om de vore tesaurustermer. Användarvokabulärer är främst avsedda för egennamn men kan även innehålla nya termer som tesaurusen bör utökas med. Användarvokabulärer öppnas med bokhylleknappen (på höger sida).

Om LexWare®

LexWare® är en språkanalysator som utgör kärnan i tillämpningar som kräver textanalys. Den bygger på omfattande kunskaper i svenska.

LexWare® lexikon

- 80000 enheter i huvudlexikonet, baserat på Nationalencyklopedins ordbok
- Basordlistor: engelska, latin, franska, tyska
- 50000 namn: människor, platser, organisationer, mm
- Ordformsrepresentation motsv. 800000 ordformer:
  - 75 grammatiska kategorier
  - 11 ordklasser
  - 48 ordbildningstyper
- Ordbildningslänkar:
  - till avledningar
  - till ordled
- Etymologi
- Stil
- Ämnesklassificering med c:a 100 ämneskategorier
- Synonymilänkar och parafraser
- Tesauriska länkar

LexWare® grammatik

- 400 ordbildningsregler för dekomposition i huvud- och baskomponent
- 500 generella frasbildningsregler samt 700 fasta frasförbindelser
- Separat körtidsrepresentation av grammatik och lexikon
- En ”lat” analysstrategi (Dura, E. 1998. Parsing Words. ISBN 91-87850-16-8)

LexWare® program

- Utveckling påbörjad 1995, registrerad varumärke 2000
- Prestanda: c:a 20000 textord per sek. med en P III 800Mhz
- Täckning: 98% med unik igenkänning
- Portabilitet: språkanalysator i standard C, tillämpningar i standard Java
- Extern utvärdering: bäst i djupindexering av Riksdagens dokument (Kristina Bäckström, Uppsala universitet)