

Uppsala universitet
Institutionen för lingvistik
Språkteknologiprogrammet

EXAMENSARBETE
Vt 2000

Marknadsundersökning och utvärdering
av indexeringsprogram
—
en delstudie inom projektet
Automatisk indexering

Kristina Bäckström

Handledare: Anna Sågvall Hein
Institutionen för lingvistik

Sammandrag

Detta examensarbete behandlar frågan huruvida automatisk eller datorstödd indexering kan vara ett alternativ till manuell indexering. Med indexering menas här tilldelandet av relevanta nyckelord till ett dokument med utgångspunkt från dokumentets innehåll.

Arbetet har utförts på Riksdagsbibliotekets sektion för indexering och registerproduktion, där ca fyra indexerare årligen manuellt indexerar över 4000 av riksdagens dokument, såsom motioner, propositioner, frågor och interpellationer.

Syftet med arbetet är att undersöka hur bra indexering av riksdagens dokument som kan åstadkommas med hjälp av dagens indexeringsprogram jämfört med manuell indexering. Arbetet har huvudsakligen bestått av två uppgifter:

- En marknadsundersökning som resulterade i kunskaper om vilka marknadsaktörer och indexeringsprogram som finns.
- En jämförande studie mellan manuell och automatisk indexering. Fyra olika programvarors indexering jämfördes med två indexerares. Programmens indexering utvärderades i termer av indexeringskonsistens, recall, precision och F-värde. Närmare 200 riksdagsdokument ingick i testmaterialet.

I examensarbetet beskrivs även indexering i allmänhet och indexeringen på Riksdagsbiblioteket i synnerhet, samt några grundläggande språkteknologiska metoder som kan användas för att skapa indexeringsprogram.

Innehållsförteckning

<i>Sammandrag</i>	3
<i>Innehållsförteckning</i>	5
1 Inledning	7
2 Bakgrund	9
2.1 Indexering	9
2.2 Indexering på Riksdagsbiblioteket	11
2.2.1 Indexeringsarbetet	11
2.2.2 Dokumenten	12
2.2.3 Tesaurusen	13
2.2.4 Lagring av dokument och information.....	14
3 Språkteknologiska indexeringsmetoder	16
3.1 Frekvensanalys	16
3.2 Grundformsigenkänning	17
3.2.1 Stamning	17
3.2.2 Lemmatisering	17
3.3 Relevanta ord	18
3.3.1 Stoppordlistor.....	18
3.3.2 Inversed Document Frequency	18
3.4 Frashantering	19
3.4.1 ”N-ordningar”	19
3.4.2 Syntaktisk analys.....	19
3.5 Utvidgningar och andra metoder	19
4 Undersökning av marknaden	21
4.1 De olika informationsvägarna	21
4.2 Företag, forskningsorganisationer och produkter	22
Oracle	22
Conexor.....	22
Institute for Information Technology, National Research Council Canada.....	23
IBM Intelligent Miner for Text.....	23
Excalibur Technologies.....	24
Verity Inc.	24
Circle Noetic Services.....	25
InXight	25
Iconovex Corporation	26
Lernout&Hauspie (L&H Mendez).....	26
GSI-Erli.....	26
The Jelem Company.....	27
Data Harmony	27
University of Edinburgh, The Language Technology Group	27
Megaputer Intelligence	28

Lingsoft	28
LexWare Labs	29
Word Smith Tools	29
SICS	30
Sunstone Systems AB	30
Autonomy.....	31
KTH, NADA.....	33
4.3 Sammanfattning – marknaden	34
5 Urval	35
6 Utvärdering av indexeringsprogram	37
6.1 Metod.....	37
6.2 Material.....	39
6.3 Den manuella indexeringen.....	41
6.4 Förutsättningar för testerna	42
6.5 Resultat	43
6.5.1 Lingsoft	43
6.5.2 Conexor.....	47
6.5.3 LexWare Labs	49
6.5.4 KTH	51
6.6 Sammanfattning av testerna	53
7 Sammanfattning – diskussion.....	57
Litteraturförteckning.....	59
Bilaga Lingsoft.....	61
Bilaga Conexor.....	69
Bilaga LexWare Labs.....	77
Bilaga KTH.....	84
Diagram 1: Följdmotioner, Lingsofts alla och indexerarnas termer	44
Diagram 2: Följdmotioner, Lingsofts tio högst rankade och indexerarnas termer	45
Diagram 3: Följdmotioner, Lingsofts filtrerade och indexerarnas termer	46
Diagram 4: Följdmotioner, Conexors alla och indexerarnas termer	47
Diagram 5: Följdmotioner, Conexors tio högst rankade och indexerarnas termer	48
Diagram 6: Följdmotioner, Conexor filtrerade och indexerarnas termer.....	49
Diagram 7: Följdmotioner, LexWare Labs alla och indexerarnas termer.....	50
Diagram 8: Följdmotioner, LexWare Labs tio högst rankade och indexerarnas termer	51
Diagram 9: Följdmotioner, KTH:s alla och indexerarnas termer	52
Diagram 10: Indexeringskonsistens, jämförelse mellan kandidaterna.....	53
Diagram 11: Recall, jämförelse mellan kandidaterna	54
Diagram 12: Precision, jämförelse mellan kandidaterna	54
Diagram 13: F-värde, jämförelse mellan kandidaterna.....	55

1 Inledning

I föreliggande rapport redovisas mitt examensarbete på Språkteknologiprogrammet vid Uppsala universitet. Det bygger i huvudsak på det arbete som jag utfört som en delstudie inom projektet Automatisk indexering, vilket bedrivs på Riksdagsbiblioteket vid sektionen för indexering och registerproduktion. Delstudien har utförts under elva månader, med början i mars 1999.

Det övergripande syftet med projektet är att undersöka möjligheterna att helt eller delvis automatisera den indexering av dokument som görs på sektionen. Indexering innebär här tilldelandet av nyckelord till ett dokument med utgångspunkt från dokumentets innehåll. Delstudien syftar till att undersöka hur bra indexering av riksdagens dokument som kan åstadkommas med hjälp av dagens indexeringsprogram jämfört med manuell indexering.

De två huvudsakliga uppgifter som utförts inom ramen för mitt arbete är följande:

- En marknadsundersökning av företag och produkter för automatisk/datorstödd indexering.
- En jämförande studie mellan fyra programvarors indexering och manuell indexering och en utvärdering av programmen.

De två huvuduppgifterna redovisas i varsitt kapitel i rapporten, kapitel fyra respektive kapitel sex. Kapitel fem sammanbinder de två kapitlen.

Kapitel två är ett bakgrundskapitel som dels beskriver vad den bibliografiska verksamheten indexering innebär och dels sätter in marknadsundersökningen och den jämförande studien i sitt sammanhang genom att beskriva hur dagens manuella indexering går till på Riksdagsbiblioteket.

Kapitel tre redogör för vissa grundläggande språkteknologiska indexeringsmetoder som är vanligt förekommande i olika indexeringsprogram. Detta kapitel underlättar förhoppningsvis läsningen av de följande kapitlen vad gäller språkteknologisk terminologi.

Marknadsundersökningen behandlas i kapitel fyra. De företag/forskningsorganisationer/produkter som hittats i genomgången av marknaden listas och beskrivs kort. De företag/forskningsorganisationer/produkter som vi haft närmare kontakt med beskrivs mer utförligt.

Kapitel fem är, som nämnts, länken mellan marknadsundersökningen och den utvärdering av indexeringsprogram som gjorts. Där nämns vilka indexeringsprogram som till slut valdes ut för att testas. Kapitel fem beskriver också de kriterier som låg till grund för urvalet.

Testerna och utvärderingen av de fyra indexeringsprogrammen återfinns i kapitel sex. Där beskrivs också den metod som använts för utvärderingen. Varje programs resultat redovisas i ett eget delkapitel och sedan följer en övergripande sammanfattning.

Rapporten avslutas med kapitel sju, som är en sammanfattning av hela rapporten, d.v.s. både marknadsundersökningen och utvärderingen av indexeringsprogrammen. Där återfinns också tankar om vad indexering är och syftar till samt eventuella framtida utvidgningar av det arbete som gjorts inom den här delstudien.

2 Bakgrund

2.1 Indexering

Det finns många definitioner av indexering och inom olika områden har ordet olika innebörd. En definition som vanligtvis används för den typ av indexering som görs på bibliotek innebär att indexering är en intellektuell bibliografisk verksamhet som syftar till att förbättra informationsåtervinning. Det som indexerats kan t.ex. vara böcker, tidskrifter, artiklar och andra typer av dokument. Indexering innebär att ett dokument tilldelas ett antal ord och fraser som beskriver dokumentets innehåll (Hellsten och Rosfelt 1999:83). Istället för ord och fraser brukar man i indexeringssammanhang använda benämningen termer. Om termerna är väldefinierade och standardiserade kallas de indexeringstermer (Benito 1993:110). De utgör då ett kontrollerat språk. Indexeringstermer kan ingå i en tesaurus, som är en typ av standardiserad ordlista som bl.a. specificerar indexeringstermernas semantiska relationer till varandra (Benito 1993:114). En indexeringsterm som återfinns i en tesaurus kallas deskriptor (Benito 1993:109).

De deskriptorer som vid indexering tilldelas dokument samlas ofta i en databas eller i ett tryckt index. För varje deskriptor finns en hänvisning till de dokument som indexerats med den. Genom att söka på en deskriptor, antingen genom ett sökprogram kopplat till databasen eller genom att slå i det tryckta indexet, kan man hitta de dokument som indexerats med deskriptorn och som således handlar om ett visst ämne. Detta är indexeringens främsta syfte - att se till att informationen i dokument blir sökbar och möjlig att återfinna med hjälp av den innehållsbeskrivning som görs med deskriptorer.

Indexering är en del av informationsvetenskapen och ett vetenskapligt ämne inom vilket det finns en mängd teorier och regler. Det finns särskilt vissa saker som en indexerare bör tänka på vid tilldelandet av deskriptorer till ett dokument:

- **Uttömmandegrad:** Indexeraren bör se till att alla relevanta ämnen som behandlas i ett dokument beskrivs med hjälp av de deskriptorer som tilldelas dokumentet. Om inte detta görs kan det leda till informationsförlust, eftersom det i princip blir omöjligt för en användare att hitta dokumentet. Uttömmandegrad talar om till vilken grad de koncept och ämnen som behandlas i ett dokument kan återfinnas med hjälp av de indexerade termerna (Benito 1993:107).
- **Specificitet:** Indexeringen ska vara så specifik som möjligt. Detta innebär att de deskriptorer som indexeraren tilldelar ett dokument på ett så nära sätt som möjligt ska beskriva dokumentets innehåll. Man bör välja deskriptorer som ligger nära dokumentets egen nivå, d.v.s. deskriptorer som motsvarar de begrepp som dokumentet avhandlar. Indexeraren bör inte använda sig av för vida begrepp, utan hålla sig på dokumentets nivå. Specificitet handlar alltså om hur pass nära indexeringstermerna kommer det begrepp man vill beskriva (Benito 1993:108).

- **Objektivitet:** Indexeraren bör använda sitt omdöme och göra en rimlig bedömning av dokumentet - hur viktigt är dokumentet i jämförelse med andra dokument i dokumentinsamlingen? Han/hon får dock inte vara subjektiv i sin bedömning av dokumentets innehåll (Lindkvist Michailaki 1999:1). Indexeringen ska hållas på en så objektiv nivå som möjligt och ska inte spegla den enskilde indexerarens speciella intressen eller åsikter.

Indexering är således en komplicerad process (Karlgrén 2000:9) som kräver mycket kunskap, både om hur indexering bör gå till och om reglerna och riktlinjerna ovan, om hur den tesaurus som används i indexeringsverksamheten är uppbyggd och fungerar samt om de ämnen som behandlas i dokumenten som indexeras.

Man strävar efter att med hjälp av ovanstående riktlinjer göra en så bra indexering som möjligt. En välgjord indexering ger hög precision då en användare söker efter information, d.v.s. användaren får vid en sökning fram bara de dokument som handlar om det användaren sökt efter (relevanta dokument). En välgjord indexering ger samtidigt en hög återvinningsgrad (recall), d.v.s. användaren är säker på att vid en sökning hitta alla de dokument som handlar om det användaren sökt efter (relevanta dokument). Men även om strävan är att uppnå både hög precision och hög återvinningsgrad är det långt ifrån alltid detta blir fallet. De två kriterierna är nämligen svårförenliga. Ju mer specifika termer som används vid indexeringen, desto högre grad av precision blir det, men desto färre av de relevanta dokumenten hittas. Alltför generella termer är också meningslösa, eftersom användaren då hittar litteratur som han/hon inte är intresserad av och som han/hon själv måste sälla bort manuellt. Vilken av de två målsättningarna (hög precision eller hög återvinningsgrad) som bör prioriteras i en indexeringsverksamhet beror på dess syfte: vid retrospektiv sökning (tillbakablickande, d.v.s. när indexeringen ska kunna användas för sökning även många år efter indexeringstillfället) är det fördelaktigare med ett system som har en hög grad av precision, medan det vid löpande sökning är önskvärt med ett system som har högre återvinningsgrad (Benito 1993:107). Om indexeraren vid indexeringstillfället vet att användaren vid söktillfället kommer att ha tillgång till den tesaurus som används i indexeringen bör indexeraren indexera med mer specifika begrepp än om användaren inte har tillgång till tesaurusen. Användaren kan nämligen med hjälp av tesaurusens specificering av deskriptorernas semantiska relationer härleda det mer generella ur det specifika, men inte tvärtom. Att indexera med generella deskriptorer innebär således en irreversibel förlust av information. Använder indexeraren t.ex. deskriptorn *rågbröd* för ett dokument, så implicerar det att dokumentet handlar om *bröd*. Ett dokument som är indexerat med deskriptorn *bröd* implicerar däremot inte att dokumentet handlar om *rågbröd*; det skulle lika gärna kunna handla om *vetebröd*.

Trots de regler och riktlinjer som finns för indexering är det svårt att uppnå enhetlighet, konsekvens. Man brukar använda termen indexeringskonsistens när man talar om enhetlighet och konsekvens mellan olika indexerare och olika indexeringstillfällen. Indexering är till stor del beroende av den kunskap som en indexerare besitter och vilka associationer han/hon därmed gör vid indexeringen av ett dokument. Kunskapen varierar från indexerare till indexerare; det är oundvikligt eftersom de är människor. Kunskap är heller aldrig något statiskt hos en indexerare; varje dag lär man sig något nytt. Att vara konsekvent i indexeringen är alltså något som är svårt för en enskild indexerare. Konsekvensen blir allt svårare att upprätthålla ju fler indexerare som arbetar med samma dokumentmängd. Eftersom indexering är en relativt tidskrävande uppgift är det ofta nödvändigt att anlita flera indexerare för att få jobbet gjort. Detta leder således till inkonsekvens i indexeringen och det leder i sin tur till att användaren vid en sökning inte kan vara säker på att hitta allt relevant material och slippa irrelevanta träffar.

För att öka enhetligheten kan en indexeringsverksamhet använda sig av en tesaurus. Denna innehåller en ändlig mängd termer och bestämmer vilka ord och fraser som är tillåtna att användas för indexering. På så sätt begränsar man indexerarnas valfrihet vad gäller ordval vid indexeringen och detta leder till större enhetlighet. En begränsning i antalet termer som får tilldelas ett dokument är ett annat sätt att öka indexeringskonsistensen. Ju färre termer indexerarna får tilldela ett dokument, desto större är sannolikheten att de överensstämmer i sin bedömning av vilka ämnen som är mest relevanta för dokumentet i fråga.

Att inte uppnå en betryggande enhetlighet är således indexeringens dilemma. Trots det visar studier att sökning i indexerat material ändå ger bättre resultat än att söka i fritext (Hjørland 1993:22, Milstead 1998:2).

2.2 Indexering på Riksdagsbiblioteket

2.2.1 Indexeringsarbetet

Den indexering som görs vid Riksdagsbiblioteket kan betecknas som ett mänskligt intellektuellt arbete som består i att göra en identifiering och beskrivning av de ämnen som ett dokument handlar om. Till sin hjälp har indexerarna en tesaurus, Riksdagens tesaurus. Indexeringsarbetet kan delas upp i olika delmoment:

- **Innehållsanalys:** Indexeraren bestämmer vad dokumentet handlar om.
- **Utkristallisering av begrepp:** Indexeraren finner de begrepp som är relevanta för dokumentets innehåll.
- **Val av deskriptorer:** Indexeraren kontrollerar att de begrepp som utkristallerats som betydelsefulla för dokumentet finns i Riksdagens tesaurus (eller bland de övriga termer som används för indexering, se nedan 2.2.3). Återfinns inte begreppen så gäller det för indexeraren att i tesaurusen hitta andra deskriptorer som kan användas istället.
- **Inmatning av deskriptorer:** De deskriptorer som slutligen valts ut för att beskriva dokumentets innehåll matas in i ett visst fält i en databas.

2.2.2 Dokumenten

De olika typer av dokument som indexeras på Riksdagsbibliotekets sektion för indexering och registerproduktion är följande:

- Propositioner
- Skrivelser
- Redogörelser
- Förslag
- Motioner
- Interpellationer
- Frågor

De olika typerna av dokument har vissa likheter, men det finns också stora skillnader mellan dem. Propositioner är t.ex. mycket välskrivna eftersom de har hanterats och förbättrats av många instanser innan de föreligger i färdigt skick. Dessutom har de ofta författats av en person som har utformande av dylika dokument som sitt yrke och som således vet hur man uttrycker sig klart och tydligt. Propositionerna är ofta omfångsrika och långa och man skulle därför kunna tro att de är extra arbetsamma att indexera. Men tack vare deras explicita struktur och genomarbetade språk så upplevs de av indexerarna som lättare att indexera än andra dokument.

Motioner kan vara av olika typer. Det finns följdmotioner som är en ledamots (eller en grupp av ledamöters) reaktion på en proposition och som således behandlar en eller flera punkter i den aktuella propositionen. Det finns allmänna motioner som kan handla om i princip vad som helst och som måste lämnas in under en särskild tid på hösten, den allmänna motionstiden. Både följdmotioner och allmänna motioner kan vara antingen partimotioner (redogör för ett partis ståndpunkt i en viss fråga), gruppmotioner (skrivna av en grupp av ledamöter) eller enskilda motioner (skrivna av en enda ledamot). Motionerna skiljer sig ofta åt beroende på vilken typ de är. Partimotionerna liknar propositionerna så till vida att de har stötts och blötts av många händer innan de föreligger i färdigt skick. Dessutom har de, liksom propositionerna, ofta skrivits med hjälp av någon person vars yrke är att skriva. Enskilda motioner och gruppmotioner kan variera i kvalitet. Vissa ledamöter är duktiga på att skriva och har kanske dessutom vänt sig till sitt partikansli för att få hjälp och råd i sitt skrivande. Andra lägger inte ned så mycket tid och möda på att få sitt språkbruk korrekt och att göra motionens struktur klar och logisk.

Även de andra dokumenttyperna har sina speciella kännetecken. Interpellationer och frågor är sällan längre än en A4-sida (oftast mycket kortare) och har inte samma struktur som en motion eller en proposition.

Totalt indexeras ca 4500 dokument per riksmöte (september-september) och den överlägset största gruppen av dokument är de allmänna motionerna.

2.2.3 Tesaurusen

Den tesaurus som används för indexeringsverksamheten på Riksdagsbiblioteket består av ca 3500 termer. Den är utarbetad för att passa riksdagens indexeringsbehov och dess ämnesområden följer sålunda i stort sett utskottens ämnesuppdelning. Eftersom riksdagen avhandlar frågor av de mest skilda slag, från barnomsorg till rymdprogram och exportpolitik, spänner tesaurusen över ett vitt spektrum av begrepp. I tesaurusen specificeras följande information:

- Deskriptorer: De termer som indexerarna använder vid indexeringen av dokument. Deskriptorerna utgör ett kontrollerat språk och är auktoritetskontrollerade, d.v.s. de är godkända av riksdagens indexeringsverksamhet som indexeringstermer.
- Icke-deskriptorer: De termer som inte ska användas för indexering. De är inte upptagna bland den auktoritetskontrollerade mängden av termer, men är t.ex. ändå vanligt förekommande i dokument och spridda i allmänspråket så att indexerarna kanske känner sig lockade att använda dem. Varje icke-deskriptor är försedd med en hänvisning till den deskriptor som bör användas istället.
- Scope Notes (SN): En del av deskriptorerna i tesaurusen har försetts med en Scope Note. Det kan vara en definition av deskriptorn, eller en förklaring som anger omfattningen av eller meningen med den och hur den används.
- Broader Term (BT): Överordnad term. Termerna i tesaurusen är ordnade i betydelsehierakier. En överordnad term har en vidare betydelse än dess underordnade termer. Deskriptorn *Alternativ medicin* är t.ex. BT till deskriptorn *Akupunktur*. *Akupunktur* är helt enkelt en typ av *Alternativ medicin*, och *Alternativ medicin* har en vidare betydelse än *Akupunktur*.
- Narrower Term (NT): Underordnad term. En underordnad term har en mer begränsad betydelse än sin överordnade term. Se Broader Term för exempel.
- Related Term (RT): Besläktad, parallell term. *Folkrörelser* och *Folkhögskolor* är exempelvis RT till varandra. De har ingen hierarkisk relation till varandra, utan bara en allmän semantisk (<http://instruct.uwo.ca/gplis/677/thesaur/main00.html>).

Tesaurusen är en indexerares främsta uppslagsbok. Han/hon använder den för att kontrollera att den deskriptor han/hon tänkte tilldela ett dokument verkligen finns upptagen i tesaurusen och för att få uppslag om vilka termer som kan användas för att indexera ett visst ämnesområde. Tesaurusen är sålunda flitigt använd vid sektionen för indexering och registerproduktion.

Tesaurusen innehåller en mängd begrepp som beskriver omvärlden, och eftersom omvärlden ständigt förändras behöver även tesaurusen förnyas, förändras och uppdateras. Den är sålunda ett levande verktyg som aldrig är ”färdigt” (Riksdagens tesaurus 1997:2, Benito 1993:105).

Förutom de deskriptorer som finns i tesaurusen används också en del andra termer i indexeringsarbetet på Riksdagsbiblioteket. Det är namn på geografiska platser, företag samt statliga verk och myndigheter. Dessa termer är inte samlade på samma konkreta vis som deskriptorerna i tesaurusen, och de är heller inte auktoritetskontrollerade på samma sätt, men på Riksdagsbibliotekets sektion för indexering och registerproduktion kallas de ändå deskriptorer

eftersom de används för indexering på samma sätt som tesaurusens deskriptorer. I rapporten benämns de fortsättningsvis namnord för att kunna skilja dem från deskriptorerna i tesaurusen.

Som nämnts ovan kan man uppnå större enhetlighet indexering emellan om man använder en tesaurus som indexeringsverktyg och om man begränsar antalet deskriptorer som får tilldelas ett dokument. Tesaurusen utgör en viktig del av indexeringsverksamheten på Riksdagsbiblioteket och man har också riktlinjer för antalet deskriptorer per dokument. Propositioner, skrivelser, förslag och redogörelser bör inte indexeras med fler än 20 deskriptorer, motioner bör inte tilldelas fler än 10 deskriptorer och interpellationer och frågor bör inte indexeras med fler än 5 deskriptorer. Dessa siffror utgör som sagt riktlinjer och får därmed överskridas i speciella fall. För de flesta dokument är dock gränserna väl tilltagna och det är sällan som t.ex. en fråga indexeras med så många som 5 deskriptorer.

2.2.4 Lagring av dokument och information

Det finns två olika typer av databaser för lagring av riksdagens dokument – dels fulltextdatabaser och dels ett sakregister som består av mer kortfattad information om varje dokument. Båda typerna av databaser finns tillgängliga för allmänheten på hemsidan <http://www.riksdagen.se/debatt/Index.asp>.

Fulltextdatabaserna innehåller, som namnet antyder, alla dokument i fulltext, d.v.s. hela dokumenten från början till slut, oavsett hur långa de är. För att söka i fulltextdatabaserna använder man sig av fritextsökning. Man kan alltså söka på vilket ord man vill, *och, så, men*, men också efter innehållsord såsom *miljöpolitik, fastighetsskatt, barnomsorg* och få träff på det. Söksystemet bygger på boolsk algebra och strängmatchning och söker endast på de tecken användaren anger i sin sökning. Systemet är således känsligt för felstavningar och kan inte hantera böjda ord. För att ändå kunna söka efter flera böjda former av samma ord kan man använda trunkering, d.v.s. man söker på t.ex. ett ords stam och anger sedan ett tecken, ett s.k. wildcard, som accepterar vilka tecken som helst.

Sakregistret är inte en fulltextdatabas, utan innehåller mer kortfattad information om varje dokument. Dokumenten är uppdelade efter yrkanden och det finns en post i databasen för varje yrkande. Varje post består av fälten Rubrik och Yrkande samt fält innehållande information om när dokumentet inkom och när det behandlades. I den post som innehåller ett dokumentets yrkande nummer ett finns fältet Nyckelord och här återfinns de termer som dokumentet indexeras med.

Det finns tre olika sätt att söka i sakregistret. Dels finns den terminalbaserade kommandosökningen där användaren själv får formulera ett sökkommando med hjälp av det fördefinierade kommandospråket. Detta används mest av den personal som sköter uppdateringen av databaserna och databasvården. Med denna metod kan man göra raffinerade sökningar på speciella fält i databaserna. Dels finns den terminalbaserade formulärsökningen. Detta kan sägas vara en förenklad kommandosökning. Sökfunktionaliteten är något begränsad och alla fält i databaserna är inte representerade med ett eget fält i sökformuläret. T.ex. finns det inget specifikt fält för att söka på just Nyckelordsfältet i databaserna. Det finns dock ett sätt att kunna söka på bara fältet Nyckelord och det åstadkoms genom att man i sökformulärets fritextfält skriver

in sökfrågan NORDS=*deskriptor*. NORDS är en förkortning för nyckelord och *deskriptor* kan vara vilken deskriptor som helst som ingår i tesaursen eller bland namnorden. Sökfrågan NORDS=*Motorvägar* ger alltså bara träff på de dokument som har indexerats med deskriptorn *Motorvägar*. Att söka på detta sätt i den terminalbaserade formulärsökningen är helt enkelt att använda ett kommando från kommandosökningen i sökformuläret. Det tredje sättet att söka i sakregistret är att söka via webben. Även här är det fråga om formulärsökning. Det är i princip samma sökformulär som i den terminalbaserade formulärsökningen och även här saknas möjlighet att söka enbart i nyckelordsfältet, om man inte skriver in sökfrågan NORDS=*deskriptor* i fritextfältet. De allra flesta som söker i sakregistret använder sig av formulärsökningen, antingen av den terminalbaserade eller den webbaserade. Det är dock få, även bland Riksdagsbibliotekets bibliotekariéer, som känner till det ”trick” som behövs för att kunna söka bara i fältet Nyckelord, och på så sätt använda sig av indexeringen i sin sökning.

De deskriptorer som tilldelas dokumenten vid indexeringen på Riksdagsbiblioteket matas alltså in i databaser och kan öka återvinningen av information vid sökning i databaserna förutsatt man känner till det speciella söksättet. En gång om året, efter att ett riksmötes dokument är färdigindexerat, plockas alla deskriptorerna fram ur databaserna och sammanställs till ett tryckt index. Detta tryckta index utgör en del av det tryckta register som ges ut till riksdagstrycket. Riksdagstrycket är en serie volymer som innehåller alla dokument som producerats på riksdagen under ett riksmöte. Det tryckta registret är en hjälp för att hitta dokument och information i denna dokumentmängd. I registrets indexdel presenteras alla deskriptorer i alfabetisk ordning. För varje deskriptor finns en hänvisning till de dokument som indexerats med den. Letar man t.ex. efter dokument som handlar om naturvård kan man titta i det tryckta indexet på deskriptorn *Naturvård* och då enkelt få en överblick över vilka dokument som, enligt indexeringen, handlar om detta ämne.

En svårighet med att söka efter information med hjälp av indexeringen är att man måste känna till deskriptorerna. Har man inte tillgång till tesaursen i sin sökning kan det vara svårt att hitta rätt. T.ex. kanske användaren söker på *Miljövård* istället för *Naturvård*, vilket för en amatör kanske kan verka vara synonymer för samma begrepp. Användaren kommer då inte att få några träffar, och definitivt inga träffar på dokument som handlar om naturvård. *Miljövård* är bara en icke-deskriptor, och det inte finns några dokument som är indexerade med det ordet. Dessutom tillhör enligt tesaursen inte den deskriptor som icke-deskriptorn *Miljövård* hänvisar till, *Miljöarbete*, och *Naturvård* samma betydelsehierarki. Tesaursen är således mycket viktig, inte bara i indexeringen utan även i sökningen efter dokument (Milstead, 1998:1).

3 Språkteknologiska indexeringsmetoder

Då indexering, som vi sett ovan, är en krävande verksamhet ligger det nära till hands att försöka automatisera den, helt eller delvis (Karlgrén 2000:9). Det finns en rad metoder som kan användas för syftet, både statistiska och lingvistiska. Här följer en beskrivning av ett urval av dem.

3.1 Frekvensanalys

En grundläggande metod för att få en uppfattning om vad ett dokument handlar om är frekvensanalys. Denna metod utarbetades först av Luhn i slutet av 50-talet (Luhn 1957, 1958, 1959) och används fortfarande i stor utsträckning. Man räknar hur många gånger orden i ett dokument förekommer och sedan rangordnar man dem efter deras frekvens - det ord som förekommer flest gånger rankas högst, det ord som förekommer näst flest gånger hamnar som nummer två i rangordningslistan o.s.v. De antaganden som denna metod bygger på är dels att ett ords närvaro eller frånvaro i ett dokument har betydelse för dess innehåll (Karlgrén 2000:16), dels att man genom att använda frekvensdata kan plocka ut ord som kan representera ett dokumentets innehåll (van Rijsbergen 1979:1). Ren frekvensanalys lämpar sig dock inte särskilt bra för automatisk indexering. Nedan presenteras några nackdelar.

Ord ur de öppna ordklasserna (substantiv, verb, adjektiv) förekommer sällan i bara en enda form i ett dokument. Oftast är orden böjda på olika sätt. Gör man en frekvensanalys på en text utan att använda en lemmatiserare (datorprogram som återför ordformer till sina grundformer, se nedan) skulle det leda till att många ord skulle dyka upp i olika skepnader som egna ord, t.ex. sju förekomster av *demokratin*, tre förekomster av *demokrati*, nio förekomster av *demokratier* och fyra förekomster av *demokratins*. Kanske skulle inte någon av ordformerna uppträda tillräckligt ofta i dokumentet för att komma tillräckligt högt upp i den resulterande frekvenslistan och därmed räknas som ett viktigt innehållsord för det aktuella dokumentet. Ett sätt att gruppera ihop böjda ord av samma stam är därför viktigt för att kunna få fram de ”rätta” viktiga orden för ett dokument.

Ett antagande som är fundamentalt inom informationssökning är att substantiv innehåller mer information än andra ord i en text (Källgrén 1984:3). Exempelvis ska deskriptorerna i en tesaurus så långt det är möjligt bara utgöras av substantiv eller substantivgrupper (nominalfraser) (<http://instruct.uwo.ca/gplis/677/thesaur/main00.html>, Hellsten och Rosfelt 1999:39). Manuell indexering beskriver således dokument endast med substantiv och nominalfraser. Frekvensanalys (i sin enklaste utformning) tar däremot ingen hänsyn till vilken ordklass ett ord har, eller huruvida ett ord förmedlar mycket information (t.ex. substantiv) eller väldigt lite information (t.ex. prepositioner) i ett lösryckt sammanhang, som ju en frekvenslista oundvikligen blir. Orden rankas bara efter hur frekvent förekommande de är i ett dokument och det är ett känt faktum att det aldrig är substantiven som är de mest frekventa orden i ett dokument, utan alltid små funktionsord med liten semantisk betydelse i det lösrycka sammanhanget (Karlgrén 2000:18).

En annan nackdel med frekvensanalys är att metoden (i sin enklaste utformning) bara ger enstaka ord, inte fraser, som resultat. Hur orden uppträder tillsammans beaktas således inte. Detta är olyckligt, eftersom fraser och kollokationer ofta är viktiga i ett indexeringssammanhang. Deskriptorn *Alkoholhaltiga drycker* är t.ex. något annat än bara deskriptorn *Drycker*, som inbegriper både *Alkoholfria drycker* och *Alkoholhaltiga drycker*. Svenskan är ju, till skillnad från t.ex. engelskan, ett språk som med lätthet kan bilda ett sammansatt ord för att skapa ett nytt begrepp. Eftersom engelskan särskriver många av de begrepp som på svenska återges med ett sammansatt ord (ihopskrivet) blir frashantering än viktigare då en engelsk text ska analyseras. Men även om svenskan ofta sätter samman två (eller flera) ord till ett nytt textord, så är frashantering ändå en viktig egenskap som frekvensanalys bör kompletteras med.

3.2 Grundformsigenkänning

Utifrån de problem som beskrevs ovan är det sålunda viktigt att kunna hitta ett ords grundform för att på så sätt få ett tillförlitligare resultat över vilka ord som verkligen är mest frekventa i en text. Det finns olika strategier för att komma fram till ett ords grundform.

3.2.1 Stamning

Stamningsalgoritmer är relativt enkla heuristiska instrument för att ta reda på ett ords grundform. Stamningen ser till att alla ord med samma stam ges samma form genom att den tar bort böjnings- och avledningsändelser från stammen. Många stamningsprogram bygger på en lista med vanligt förekommande ändelser. När ett ord presenteras för programmet letar det reda på om någon av ändelserna finns i ordet. Om så är fallet tas ändelsen bort, förutsatt att vissa villkor är uppfyllda (såsom att den orddel som är kvar efter borttagandet består av fler än X bokstäver och att minst en av de X bokstäverna är en vokal, o.s.v.) (Sparck Jones och Willet 1997:306). Denna metod är relativt lätt att implementera, men bygger på antagandet att om två ord har samma stam så refererar de till samma begrepp (och samma betydelse) (van Rijsbergen 1979:12) och detta är att generalisera för mycket. Orden *motion* (formellt förslag i beslutande församling (NEO 1996)) och *motion* (kroppsrörelser som främjar hälsan och välbefinnandet (NEO 1996)) har samma stam (*motion*), men borde inte sammanföras till samma begrepp. Homonymer med olika semantisk betydelse hanteras således inte på ett tillfredsställande sätt av en enkel stamningsalgoritm.

3.2.2 Lemmatisering

Ibland används lemmatiserare och stamningsprogram synonymt, och de har onekligen vissa likheter. De arbetar med samma uppgift, nämligen att sammanföra olika ordformer som tillhör samma lexikonord. Enligt Sproat (Sproat 1992:7) är lemmatisering uppgiften att hitta ett ords lexikonform utifrån den form av ordet som faktiskt påträffas i en text. En lemmatiserare förväntar man sig lite mer av än man gör av ett stamningsprogram. En lemmatiserare ska kunna identifiera ett ord, snarare än att gissa sig till dess stam. En lemmatiserare ska också

ange den grammatiska form som ordet i dokumentet har. Vissa lemmatiserare innehåller också en modul som delar upp inputordet i morfem (morfemanalys), men detta är inte någon nödvändig egenskap som programmet måste ha för att betecknas som en lemmatiserare (Dura 1998:43).

3.3 Relevanta ord

Att hitta de ord som bäst representerar ett dokumentets innehåll är indexeringens kärna. Att ett indexeringsprogram kan göra bedömningar av vilka ord som är relevanta för dokumentet i fråga är därför av största vikt.

3.3.1 Stoppordlistor

För att komma tillrätta med problemet att det alltid är små, betydelsefattiga funktionsord som är de mest frekventa orden i en text (och därmed de ord som rankas högst i en frekvensanalys) kan man använda sig av stoppordlistor. En stoppordlista är en lista som innehåller vanligt förekommande ord som med stor sannolikhet inte är användbara som sökord (Sparck Jones och Willet 1997:306) (eller som indexeringstermer). En sådan stoppordlista kan användas för att filtrera frekvensanalysens output och då plocka bort sådana ord vars irrelevans går att förutsäga. Sådana ord är s.k. funktionsord, exempelvis prepositioner och konjunktioner, men kan även vara ord som är extremt vanliga inom ett visst ämnesområde. Ordet *demokrati* skulle troligen kunna tillhöra en stoppordlista vid automatisk indexering av riksdagstrycket, eftersom det är ett ord som används så mycket och i så många sammanhang att dess semantiska värde minskas. En stoppordlista kan konstrueras antingen manuellt eller maskinellt.

3.3.2 Inversed Document Frequency

Med hjälp av måttet Inversed Document Frequency (IDF) kan man hitta de ord som är specifika för ett visst dokument. Man tittar då på hur vanlig en term är i dokument X jämfört med i alla andra dokument i en dokumentssamling (Karlgrén 2000:18). Om termen är ovanligt frekvent i dokument X jämfört med i alla andra dokument så tas detta som en indikation på att termen är särskilt relevant för dokument X och således bra beskriver dess innehåll. Inversed Document Frequency räknas ut på följande sätt:

$$IDF = \frac{\text{totalt antal dokument i dokumentssamlingen}}{\text{totalt antal dokument som innehåller termen}}$$

3.4 Frashantering

Som beskrevs ovan är frashantering en viktig egenskap hos ett program som ska användas för sökning eller för indexering. Vi såg också att ren frekvensanalys av en text inte hanterar fraser. För att hitta fraser i en text kan man använda sig av olika metoder.

3.4.1 "N-ordningar"

Liksom man kan räkna ut frekvensen för varje enskilt ord i en text kan man även räkna ut frekvensen för "flerordningar". Man grupperar ihop alla ord i dokumentet två och två (ord1 ord2, ord2 ord3, ord3 ord4, o.s.v.) och ser sedan hur ofta en viss konstellation (t.ex. ord2 ord3) dyker upp i texten. Likadant kan man gruppera ihop ord tre och tre eller fyra och fyra för att hitta längre fraser. Men precis som vid en frekvensanalys av ett ord i taget får man även här en stor mängd irrelevant utdata och det är inte alls troligt att de fraser som är mest innehållsrika och därför mest passande som indexeringstermer hamnar överst i en dylik frekvensrankning.

3.4.2 Syntaktisk analys

Genom att göra en syntaktisk analys kan man komma åt strukturen i meningar och därmed hitta t.ex. nominalfraser. Ofta är den syntaktiska analys som görs i indexeringstillämpningar av typen "shallow parsing", d.v.s. man gör en väldigt ytlig analys, som syftar till att ta reda på de olika konstituenterna i en mening och på så sätt få fram nominalfraserna, som är de mest intressanta i ett indexeringsperspektiv.

3.5 Utvidgningar och andra metoder

Med de metoder som har redovisats här kan man komma en bit på väg mot automatisk indexering. Man kan ur ett dokument få fram substantiv och nominalfraser som är statistiskt sett relevanta för det. En sådan typ av indexering kan säkert vara tillräcklig för många indexeringstillämpningar.

Ytterligare ett sätt att förbättra utplockningen av relevanta ord ur ett dokument är att beakta dokumentets diskurs. En text är ofta uppbyggd utifrån den klassiska uppdelningen inledning, avhandling, avslutning. Inledningen och avslutningen innehåller ofta dokumentets innehåll i komprimerad form och där återfinns ofta ord som är viktiga i dokumentet. Genom att beakta detta fenomen och vikta (ge större vikt, betydelse åt) de ord som förekommer i dessa passager kan man, tillsammans med användningen av övriga metoder ovan, få fram de termer som med stor sannolikhet är viktiga för dokumentet i fråga.

Hittills har vi dock bara behandlat metoder som åstadkommer indexering med de ord som förekommer i ett dokument. Vill man åstadkomma en indexering med termer ur t.ex. en tesaurus krävs andra metoder. Man kan t.ex. utgå från det indexeringsresultat man får av metoderna beskrivna ovan och sedan tillämpa regler som säger att om term X förekommer minst Y antal gånger i det aktuella dokumentet och alltid i term Z:s närhet så ska deskriptor W användas. Att sammanställa regler för en hel tesaurus deskriptormängd är dock ett ansevärt arbete. Men det har visats att indexeringssystem som konstruerats på detta sätt har åstadkommit bra indexeringsresultat (Subject Indexing: Principles and Practices in the 90's. 1995, www.dataharmony.com/do.html).

Indexering med en fördefinierad mängd termer kallas också Text Categorization (<http://www.ltg.ed.ac.uk/software/tcr>). De metoder som används för detta är ofta baserade på maskininlärning och bland dem kan nämnas Neural Networks (<http://www.emsl.pnl.gov:2080/proj/neuron/neural/what.html>), Genetic Algorithms (<http://www.crcpress.com/index.htm?catalog/2529>), The Vector Space Model (Salton et al. 1975) och Bayesian Networks (Turtle och Croft 1990).

4 Undersökning av marknaden

Marknadsundersökningen syftade till att hitta så många företag och produkter som möjligt av dem som finns på marknaden. Internet har använts i hög grad och har lett till många intressanta fynd och kontakter. Dessutom har andra informationsvägar utnyttjats, såsom mässor, seminarier, kontakter kontakter o.s.v. Totalt har 22 marknadsaktörer hittats. Längre fram i detta kapitel ges kort information om dem avseende bl.a. deras produkter och huruvida produkten hanterar svenska eller inte.

4.1 De olika informationsvägarna

Genomgången av Internet gjordes i två etapper. Den första blev en generell genomsökning som gav en stor del av marknaden. Den visade också att fler, kompletterande sökningar behövde göras och således gjordes senare en andra genomgång med något annorlunda sökfrågor och andra sökmotorer.

De första sökningarna inriktades i första hand mot företag som förväntades vara av intresse för projektet Automatisk indexering. För att hitta företagen gjordes enkla sökningar framför allt via sökmotorn AltaVista på Internet. De sökfrågor som användes var t.ex. "automatic indexing", "text analysis" och "content analysis" samt deras svenska motsvarigheter. En av träffarna, hemsidan MLIS (<http://guagua.echo.lu/mlis/>), visade sig vara en källa till mycket information. MLIS är ett treårigt gemenskapsprogram för det flerspråkiga informationssamhället och finansieras av EU. På MLIS hemsida återfinns information om språkteknologiska företag. De presenteras kort med namn, målsättning, produktnamn och användningsområde, adress och eventuell hemsida- och e-postadress. Efter att ha gått igenom den information som fanns på MLIS och sorterat fram sådana företag som sade sig arbeta med automatisk indexering fortsatte sökandet efter dessa på Internet. Vissa gick lätt att hitta, andra hade ingen hemsida och åter andra visade sig vara återvändsgränder. I detta skede togs ingen hänsyn till i vilket land företagen var belägna och inte heller till om deras produkt(er) uttryckligen kunde eller inte kunde hantera svenska. Huvudsaken var att få tag på så många företag och produkter som möjligt.

I sökandet efter företag/forskningsorganisationer/produkter på Internet hittades en mängd information, bl.a. företagspresentationer, White Papers¹ och forskningsrapporter, som lästes igenom och sammanställdes. Efter detta steg togs en första kontakt med många av företagen, ofta via e-post.

Efter den första genomgången av marknaden gjordes kompletterande sökningar. Vid dessa sökningar användes fler sökmotorer såsom AltaVista, Spray Sök, Yahoo, Excite, MSN Search och metasöktjänsten Copernic. Sökfrågorna var bl.a. "document categoriz(s)ation", "text categoriz(s)ation", "information retrieval", "knowledge management" och deras svenska motsvarigheter.

¹ Ett företags presentation av sin produkt, framställd på ett någorlunda vetenskapligt och objektiva sätt.

Mässan Information Management 99 med Arkiv och Dokument hölls i maj 1999 på Sollentunamässan utanför Stockholm. Där fanns en mängd utställare, varav vissa var av intresse för projektet Automatisk indexering. De kontakter som togs där ledde i några fall till uppföljande företags- och produktpresentationer på så sätt att företagen kom till Riksdagsbiblioteket för att berätta om sig och sina produkter.

Vi har även kommit i kontakt med företag på andra sätt, t.ex. genom att företagen själva tagit kontakt med Riksdagen eller att någon person i indexeringssektionens närhet har förmedlat någon kontakt.

4.2 Företag, forskningsorganisationer och produkter

Oracle

Oracle, som är ett stort företag främst specialiserat på databaser, har utvecklat en produkt som heter ConText. Den kan bl.a. åstadkomma indexering och summering av ett dokument. Förutom att göra statistiska frekvensanalyser av ord beaktar programmet även ordens inbördes förhållanden, deras olika betydelser samt deras position både i meningen och i stycket. Med dessa metoder ”destilleras” betydelsen hos ett dokument fram.

ConText hanterar bara engelska än så länge, men detta är en produkt värd att bevaka, då det från Oracles sida finns idéer om att lokalisera (d.v.s. anpassa programvaran till ett nytt språk) den till svenska. Chalmers Tekniska Högskola i Göteborg och SICS (Swedish Institute of Computer Science) har fått förfrågningar om att vara med i ett sådant lokaliseringsarbete. Då projektet Automatisk indexering visade intresse för ConText erbjöds även Riksdagsbiblioteket att vara med i ett dylikt arbete. Enligt uppgift från Oracle skulle det kosta Riksdagsbiblioteket två-fyra miljoner kronor och ta ca två månår.

<http://www.oracle.com/oramag/oracle/97-Sep/litman.html>

Conexor

Conexor är ett finskt språkteknologiskt företag som, i likhet med företaget Lingsoft, har sina rötter i Helsingfors universitet. Conexor arbetar med välkända lingvistiska metoder, t.ex. ordanalys och begränsad grammatisk analys.

Conexor har inte någon färdig produkt för indexering, men de har en rad program som de kan bygga ihop och anpassa efter kundens önskemål. Deras program hanterar svenska.

<http://www.conexor.fi>

Institute for Information Technology, National Research Council Canada

Vid Institute for Information Technology, National Research Council Canada, har en programvara som heter Extractor utvecklats. Programmet plockar ut de ord ur ett dokument som enligt dess algoritmer är särskilt betecknande för innehållet.

Extractor hanterar endast franska och engelska i nuläget. Programmets upphovsman Peter Turney är dock inte negativ till att lokalisera produkten till svenska, bara någon svenskkunnig person är villig att göra det. Enligt Peter Turney skulle en dylik lokalisering ta ca fyra månader. Han baserar detta antagande på att tidigare implementationer av andra indoeuropeiska språk tagit så lång tid.

<http://extractor.iit.nrc.ca/>

IBM Intelligent Miner for Text

IBM har utvecklat en produktsvit som de kallar för Intelligent Miner for Text. Den har funnits i två-tre år. Produktsviten innehåller delar som kan identifiera det språk ett dokument är skrivet på, plocka ut innehållsmässigt viktiga ord ur ett dokument, klustra dokument, d.v.s. dela in en samling dokument i olika grupper utifrån t.ex. deras innehåll, samt kategorisera, d.v.s. placera dokument i fördefinierade kategorier enligt deras innehåll. Programmen i Intelligent Miner for Text fungerar bättre ju mer material man tränar dem med.

Text Analysis Tools är en delmängd av produktsviten Intelligent Miner for Text. Det är en verktygslåda för språkhantering innehållande ett knippe av olika program. Den mest grundläggande funktionen är Feature Extraction som plockar ut ord av olika slag ur en text. Man kan t.ex. välja att få se Names, Terms, Words, Organizations, Persons, Places. Feature Extraction är språkspecifik och hanterar i dagsläget inte svenska. IBM:s uppfattning är att det skulle ta ca fyra månader att implementera ett nytt språk i programmet.

Resultaten av Feature Extraction används i övriga delar som ingår i Text Analysis Tools, nämligen Clustering och Categorization, och således kan Feature Extraction anses som den grundläggande delen i programsviten. Enligt IBM är Feature Extraction väldigt viktig - fungerar Feature Extraction bra så erhåller man också bra resultat från Clustering och Categorization.

Clustering och Categorization skiljer sig, enkelt uttryckt, genom att Clustering själv hittar på kategorier för att sortera upp en dokumentmängd, medan Categorization sorterar upp en dokumentmängd i de kategorier som en användare har bestämt ska finnas.

Programmen i Intelligent Miner for Text använder sig av något som kallas för IQ, Information Quotient. Ju mer relevant en term är, desto högre IQ får den. Termens IQ bestäms av många saker – termens plats i dokumentet, frekvens, typ av ord o.s.v.

Kategoriseringsdelen av programsviten är den mest intressanta för Riksdagsbibliotekets del. Den bygger dock, som nämnts, på ordutplockningsdelen som är språkspecifik och som i dagsläget bara hanterar engelska (och i viss mån franska, tyska, spanska och italienska). Detta är

dock ett program som bör undersökas i framtiden för att se om lokaliseringar till andra språk görs.

www.software.ibm.com/data/iminer/fortext/

Excalibur Technologies

Excalibur Technologies är ett amerikanskt företag som har kontor runtom i världen. De har en produkt som heter Excalibur RetrievalWare som snarare är en sökmotor än en programvara för indexering. Den kan söka i fulltext och använder sig för svenskans del av Lingsofts morfologimodul² för att kunna hantera böjda ord i sökfrågor och därmed få bättre sökningar. Efter att vi presenterat våra problem och önskemål kring automatisk indexering på Riksdagsbiblioteket gjorde den representant vi talat med den bedömningen att Excalibur RetrievalWare nog inte var rätt produkt för den här tillämpningen.

<http://www.excalib.com/>

Verity Inc.

Verity är ett amerikanskt företag med huvudkontor i Californien. Enligt deras hemsida är deras affärsidé följande: "Verity's mission is to lead the market for knowledge retrieval solutions by turning unstructured text-intensive information into usable and shareable knowledge." Veritys produkter säljs i Sverige av företaget Nocom i Uppsala.

Nocom är ett publikt företag sedan januari 1999. Deras verksamhetsidé är "att hjälpa nordiska företag och organisationer att utveckla sin verksamhet genom att erbjuda innovativa programvaror och tjänster för effektivare informationshantering".

Företaget arbetar i projekt mot både kunder och systempartner och är en slags mellanhand mellan dessa. De systempartner de har är StoryServer, Netscape, Verity, DOCSopen (uppköpt av Fulchrum), CyberDocs, LOTSONline och SOS.

Både Verity och Fulchrum har enligt Nocom likvärdiga produkter för informationshantering, men Nocom har valt att rekommendera Veritys produkter på grund av att de kan dessa produkter bättre.

De produkter som skulle kunna vara av intresse för projektet Automatisk indexering är dels Information Server 97 och Knowledge Organizer. Information Server 97 kan göra fulltext-indexering av dokument och dessutom spara dessa och en s.k. collection innehållande författare, titel, storlek på dokumentet, summary och sökväg till dokumentet för att åstadkomma snabbare och bättre sökning. Knowledge Organizer kan kategorisera dokument, men kräver en

² Datorprogram som analyserar den interna strukturen i ord. Programmet återför ord till deras grundform och bestämmer deras böjningsform.

hel del manuellt arbete, särskilt i ett uppstartningsskede.

Tyvärr kändes ingen av produkterna särskilt relevant för projektet Automatisk indexering.

<http://www.verity.com>

<http://www.nocom.se>

Circle Noetic Services

Circle Noetic Services är ett amerikanskt företag som grundats av forskare i lingvistik och datavetenskap vid MIT (Massachusetts Institute of Technology), USA. De har utvecklat en produkt som heter WordFan. Den är ett knippe av program och kallas också Natural Language Processing Toolbox. Den kan bl.a. användas för att åstadkomma morfologiska analyser av ord³ samt att klassificera dokument utifrån innehåll. WordFan är inte färdigutvecklad för svenskans del. Det finns dock planer på en svensk version.

<http://www.CircleNoetics.com>

InXight

InXight, med säte i Palo Alto i Californien, är ett spinnoff-företag till Xerox forskningscentrum. De sysslar med Knowledge Management och har ett antal program som tillsammans kallas LinguistX. Med LinguistX kan t.ex. frasanalys⁴ och morfologisk analys åstadkommas för att hitta och plocka fram viktig information ur ett dokument. Den morfologiska analysen baseras på samma bakomliggande teori (tvånivåmorfologi) som företagen Lingsoft och Conexor använder sig av.

Många andra företag som producerar och marknadsför språkteknologiska produkter använder sig av InXights program och inkorporerar dem i sina egna produkter. T.ex. finns InXights teknologi i Oracles produkt ConText.

<http://www.inxight.com/>

http://www.inxight.com/Products/Developer/AD_Platform.html

³ Återföra ord i en text till deras grundform och bestämma böjningsform.

⁴ Fras: ett eller flera ord som tillsammans bildar en satsdel (Crystal 1994:90) och som bildar en semantisk enhet.

Iconovex Corporation

Iconovex är ett dotterföretag till Innovex och finns i USA. Enligt Iconovex företagspresentation utvecklar och marknadsför de mjukvara för automatisk indexering. Deras mest intressanta produkt för Riksdagsbiblioteket heter Syntactica och verkar vara en mycket komplicerad och kunskapsintensiv produkt som använder sig av både morfologi⁵, syntax⁶ och semantik⁷ för att tolka innehållet i ett dokument och skapa ett ämnesindex. På grund av detta faktum är tyvärr Syntactica språkspecifik för engelska och hanterar således inte svenska.

<http://www.iconovex.com>

Lernout&Hauspie (L&H Mendez)

Det belgiska Lernout&Hauspie är ett stort företag inom språkteknologi och har på senare år skaffat sig kompetens även i skandinaviska språk. Det tidigare svenska språkteknologiföretaget WordWorks i Göteborg ingår numera i Lernout&Hauspie och de samarbetar även med det relativt nystartade NST (Nordisk Språkteknologi). NST har sitt säte i norska Voss och använder sig av Lernout&Hauspies teknik för att skapa språkteknologiska produkter för den nordiska marknaden. Hittills koncentrerar sig NST på talteknologi.

Lernout&Hauspie har en produkt som heter IntelliScope® Topic Identifier. Enligt produktbeskrivningen kan den hitta intressanta och viktiga ämnen i ett dokument. I dagsläget hanterar den inte svenska.

<http://www.lhs.com>

<http://www.lhs.com/tech/icm/retrieval/toolkit/dc.asp>

GSI-Erli

I Europaparlamentet gjordes 1995 en studie angående möjligheterna till automatisk indexering av parlamentets dokument (Bureau van Dijk 1995). Ett av de företag som ingick i undersökningen var GSI-Erli. På Internet finns en sida om deras produkt AlethIR men denna sida har inte uppdaterats sedan 1994.

⁵ Morfologi beskriver ordens inre struktur, d.v.s. de minsta betydelsebärande elementens kombinationer till ord (Ljung/Ohlander 1982:24).

⁶ Syntax beskriver hur ord kombineras till större enheter som fraser, satser och meningar (Dahl 1982:18).

⁷ Semantik beskriver ordens och satsers betydelse (Ljung/Ohlander 1982:25).

The Jelem Company

The Jelem Company är ett amerikanskt företag som har specialiserat sig på indexering och tesaurusutveckling. Kunderna består till största delen av databasleverantörer och olika myndigheter. Företaget har även kompetens inom området datorstödd indexering (Machine Aided Indexing). De har dock ingen egen produkt som de marknadsför, utan brukar i sina projekt använda sig av företaget Data Harmonys produkt (se nedan).

<http://www.jelem.com>

Data Harmony

Data Harmony är ett amerikanskt företag och ett dotterbolag till Access Innovations. De har producerat ett indexeringsprogram som heter Machine Aided Indexer (M.A.I.). Det är ett program som från början är byggt för att indexera (till skillnad från övriga företags produkter, som egentligen är konstruerade för andra syften men som möjligtvis skulle kunna användas för indexering). Det bygger på att ett regelverk, tillsammans med en slags frekvensanalys över hur ofta varje regel används i ett dokument, föreslår termer för en mänsklig indexerare som sedan godkänner eller avvisar den föreslagna termen. Enligt Data Harmony har produkten använts för att indexera både franska, holländska, tyska och ryska texter (förutom engelska). Dock skulle det nog krävas en hel del anpassningar av produkten innan den kunde fungera för svenska, eftersom den innehåller och använder sig av en hel del lingvistisk kunskap. Dessutom måste reglerna som används i programmet anpassas och skrivas om för varje tillämpning. Men det är ett mycket intressant alternativ som är väl värt att beakta.

<http://www.dataharmony.com>

University of Edinburgh, The Language Technology Group

The Language Technology Group vid University of Edinburgh har under professor Marc Moens ledning utvecklat en produkt som kallas LT TCR (akronym för Language Technology Text Categorization and Routing). Produkten är i sig intressant och är tänkt att användas som ett datorstöd vid indexering. Det var dock några år sedan detta arbete gjordes (1994-95) och enligt hemsidan verkar det inte ha hänt så mycket sedan dess. Frågan är om LT TCR verkligen var (är) en kommersiell produkt eller bara ett forskningsprojekt. Troligen hanterar den bara engelska. Projektet Automatisk indexering har kontaktat University of Edinburgh, men har ej fått något svar.

www.ltg.ed.ac.uk/software/tcr

www.hcrc.ed.ac.uk/AnnualReport95/Text/hi4-www.html

Megaputer Intelligence

Megaputer Intelligence är ett företag med säte i Bloomington, Indiana, USA. De har en produkt som heter TextAnalyst och med denna kan man ”destillera” fram betydelsen hos en text i form av ett semantiskt nätverk, d.v.s. en grafisk bild av de viktigaste begreppen i texten och förhållandena mellan dessa begrepp viktade utifrån deras relativa betydelse. Dessutom kan TextAnalyst åstadkomma summeringar av texter och användas för att söka semantiskt (betydelsemässigt) i en text.

Produkten verkar mycket intressant, men med tanke på hur mycket lingvistisk kunskap som finns i den torde den vara språkspecifik för engelska (som den är utvecklad för) och det skulle i så fall innebära mycket arbete att lokalisera den till svenska.

<http://www.megaputer.com>

Lingsoft

Lingsoft är ett finskt företag som har sina rötter i Helsingfors universitet. En av grundarna heter Kimmo Koskenniemi och hans avhandling om tvånivåmorfologi (Koskenniemi, 1983) har varit av stor betydelse inom datorlingvistisk forskning och ligger till grund för många språkteknologiska system.

Lingsoft utvecklar mjukvara för databehandling av språk, särskilt av språken i Europa. En av deras produkter heter SWETWOL och den åstadkommer morfologisk analys av svenska ord. Den är ett generellt redskap som analyserar ord och delar upp dem i en grundform och en beskrivning av ordets böjning. Analysen bygger på en omfattande ordlista samt en beskrivning av varje ords böjnings- och sammansättningsmöjligheter.

Lingsoft är ett relativt gammalt företag inom den här branschen (grundades 1986) och deras produkter och tjänster har anlitats av ett antal företag och organisationer för en mängd applikationer. Viktigast just nu är kanske att Lingsoft har levererat den svenska grammatikkontrollen i Microsoft Office2000.

Projektet Automatisk indexering har haft omfattande kontakter med Lingsoft och presenterat indexeringsverksamhetens behov och önskemål. Lingsofts förslag var att använda en produkt som kan plocka ut nominalfraser ur en text och att sedan göra en statistisk rankning som plockar fram de mest frekventa nominalfraserna. De har sedan konstruerat ett icke-riktsdagsspecifikt indexeringsprogram som vi har haft tillfälle att utvärdera.

<http://www.lingsoft.fi>

LexWare Labs

LexWare Labs är ett språkteknologiskt företag med säte i Göteborg. Deras produkt har namnet LexWare och den kan användas för informationssökning och informationsextraktion. Kärnan i produkten analyserar ord och fraser och baseras på resultatet av en avhandling, "Parsing Words", som lades fram av LexWare Labs grundare, Elzbieta Dura, vid Göteborgs universitet 1998 (Dura, 1998).

LexWare är ett kunskapsbaserat system som använder sig av ett stort lexikon strukturerat som en databas innehållande lexikala enheter och lexikala regler för ordbildning och sammansättning. Det är därför mycket tillförlitligt vad gäller analyser av ord, d.v.s. morfologisk analys. Med hjälp av ytterligare språkteknologiska instrument har företaget utvecklat en prototyp som kan användas för att indexera texter med deskriptorer ur Riksdagens tesaurus. De har länkat deskriptorerna i tesaurusen med LexWares lexikala enheter. LexWare identifierar termerna i en löpande text och beräknar potentiella deskriptorer utifrån termernas förekomster. Relaterade termer betraktas i denna beräkning som förstärkande faktorer.

<http://www.lexwarelabs.com>

Word Smith Tools

Word Smith Tools är ett program som har utvecklats av Mike Scott i Liverpool, England. Det är främst skapat för att i studiesyfte analysera korpora (textsamlingar). Programmet består av flera moduler, med vilka man kan bearbeta stora textmassor och få fram olika statistiska resultat.

En av modulerna är KeyWords och den kan användas för att plocka ut nyckelord ur en text. Andra moduler ur Word Smith Tools måste dock användas innan KeyWords slutligen kan indexera ett dokument. Först måste en referenskorpus skapas genom att man samlar ihop en stor mängd text och slår ihop allt detta till ett enda stort dokument. Referenskorpusen ska representera "normalt språkbruk" och ska innehålla ord av alla typer. Sedan låter man en modul ur Word Smith Tools skapa en frekvenslista över de i referenskorpusen ingående orden: Orden räknas i referenskorpusen och den resulterande frekvenslistan upptar orden tillsammans med en siffra som anger hur många gånger ordet i fråga förekom i referenskorpusen. Sedan görs samma sak med det dokument man vill indexera. Själva nyckelordsutplockningen, indexeringen, baseras på en jämförelse av hur många gånger ord X förekommer i referenskorpusen och i det dokument man har för avsikt att indexera. Om ord X procentuellt sett förekommer mycket oftare i indexeringsdokumentet än i referenskorpusen, så antas detta vara en indikation på att ord X är ett framträdande, och således innehållsmässigt viktigt, ord i indexeringsdokumentet och därför kan tjäna som ett nyckelord för det.

Ett problem med Word Smith Tools är att det beaktar ALLA ord som ingår i ett dokument. Det har ingen funktion för att lemmatisera för att beakta ord med samma stam som förekommer av samma ord, utan varje böjningsform blir ett eget ord i frekvenslistan.

En annan nackdel med Word Smith Tools är att det inte finns någon möjlighet att sortera ut vissa grupper av ord, t.ex. substantiv. I en indexeringsituation är andra ordklasser såsom verb,

adverb och prepositioner irrelevanta eftersom endast substantiv och substantivfraser används som deskriptorer då de tycks innehålla mer information än andra ord i en text (Källgren, 1984:3). Word Smith Tools gör dock ingen åtskillnad på de olika ordklasserna utan presenterar helt enkelt alla ord i rangordning efter hur viktiga de är som representanter för dokumentet i fråga.

Word Smith Tools har heller ingen funktion för att hantera fraser utan betraktar allt som åtskiljs med mellanslag som olika semantiska enheter. Frashantering är mycket viktigt för ett språk som engelska där många sammansättningar är särskrivna, men även för ett språk som svenska, där sammansättningar (oftast) är ihopskrivna, är frashantering av stor vikt. Termen "Militär utbildning" beskriver t.ex. något annat än den enkla termen "Utbildning".

<http://www.liv.ac.uk/ms2928/wordsmith/>

SICS

Vid SICS (Swedish Institute of Computer Science), Kista, finns en språkteknologigrupp. Ett av deras projekt heter Digitala bibliotek. En sammanfattning av det projektet lyder som följer:

"Kunskapsintensiva organisationer ställer speciella krav på bibliotek. Detta projekt avser studera informationsbehov hos kunskapsarbetare i kunskapsintensiva organisationer; utveckla och utprova teknik för automatisk informationsspridning inom en organisation; bygga språkteknologiskt baserade verktyg för att stödja dagens informationsspecialister att organisera och inhämta information inom en organisation, och utvärdera de resulterande verktygen baserat på funna informationsbehov. "

Bland de uppgifter som planeras av SICS i projektet Digitala bibliotek och som är av direkt intresse för projektet Automatisk indexering på Riksdagsbiblioteket kan nämnas följande:

"Utveckling av indexeringsstödsystem för professionell indexering av svenska textsamlingar, givet givna terminologiska direktiv i form av indextermdatabaser. Speciell uppmärksamhet kommer att riktas på textsamlingars nödvändiga föränderlighet över tiden."

SICS har i dagsläget inget indexeringsprogram. De använder sig av Lingsofts produkt SWETWOL för att skapa olika prototyper för experimentverksamhet.

<http://www.sics.se>

Sunstone Systems AB

Sunstone Systems AB är ett Stockholmsföretag med inriktning på dokumenthantering och fritextsökning. De producerar inte själva några språkteknologiska program, men de är återförsäljare av de produkter som det amerikanska företaget DataWare skapar. Dessutom gör

de en del anpassningar av dessa produkter för den svenska marknaden. De har två produkter, InQuery och Categorization Server, som kan vara av intresse för projektet Automatisk indexering.

InQuery är en fritextindexeringsprodukt som är tänkt att användas som en robot som surfar hem hemsidor vars adresser användaren angivit. Programmet indexerar sedan sidorna och gör dem sökbara. InQuery hittar inte bara alla ord som finns på hemsidan utan även koncept, företagsnamn och personnamn. InQuery kan användas för att söka i sådant material som den redan har indexerat och vid en sökning plockar den inte bara fram de mest relevanta hemsidorna för den aktuella sökningen, utan anger även vilka koncept, företagsnamn och personnamn som man kan söka vidare på eller begränsa sin sökning med. Användaren kan också låta programmet arbeta mot egna koncept och specificera regler för vad programmet är tvunget att hitta i ett dokument för att tilldela det ett visst koncept. Ett exempel som belyser detta är att användaren kan ange en regel som säger att om det kommer ett eller två eller flera ord som börjar med stor bokstav och om det (eller de) följs av AB, så ska det uppfattas som ett företagsnamn och således anges som ett relevant företagsnamn.

Categorization Server är en produkt som ursprungligen utvecklades för det amerikanska patentverket. Dess funktion påminde om Autonomys Categorizer (se nedan) och IBM:s Intelligent Miner for Text. Användaren bestämmer att man behöver X antal kategorier att dela in en dokumentmängd i. Man låter sedan programmet läsa ett antal dokument för varje kategori. Dessa dokument måste användaren ha valt ut som särskilt representativa för den aktuella kategorin. Programmet lär sig i denna fas hur ett dokument ska "se ut" för att tillhöra en viss kategori. Sedan kan man mata programmet med hur många dokument man vill och om användaren tycker att programmet kategoriserar felaktigt går detta att korrigera manuellt genom att användaren "flyttar" dokumentet till en annan kategori. Programmet "lär då om" och det blir smartare och bättre ju mer det får hålla på att kategorisera dokument och ju mer användaren "hjälp till". Användaren har också möjlighet att specificera vissa nyckelord för vissa kategorier, finns t. ex. orden "bollträ", "boll" och "cheerleader" med i ett dokument så ska det hamna i kategorin "baseball". Ett dokument kan hamna under flera kategorier och om programmet råkar på ett dokument som det inte kan klassificera på grund av att det inte tycker att det finns någon kategori som är lämplig för just det dokumentet så hamnar det i en slags "överbliven hög" och väntar på klassificering av en människa. Denna produkt finns ännu inte för svenska.

<http://www.sunstone.se>

Autonomy

Autonomy är ett internationellt företag som har ett lokalkontor i Stockholm. I Sverige säljs deras produkt också av företaget Information Highway.

Autonomys produktsvit är egentligen ett Knowledge Management system, men vissa delar av det är av intresse för indexeringsverksamheten vid Riksdagsbiblioteket.

Autonomys produkter är resultatet av mångårig forskning inom neurala nät och mönstermatchning vid universitetet i Cambridge, England.

Företagets uttrycker sina mål och sin verksamhetsidé på följande sätt: "Autonomy's architecture combines innovative high-performance pattern-matching algorithms with sophisticated contextual analysis and concept extraction to automate the categorization and cross-referencing of information, improve the efficiency of information retrieval and enable the dynamic personalization of digital content."

Styrkan hos produkterna ligger i användningen av mönstermatchningsalgoritmer. Den bakomliggande teorin bygger på Claude Shannons informationsteoriprinciper, Bayes sannolikhetslära och de senaste rönen inom neurala nät. Enligt Autonomy kan denna teknik analysera en text och identifiera nyckelbegreppen i texten därför att tekniken förstår hur termernas frekvens och inbördes förhållande hänger ihop med betydelse. Eftersom tekniken inte bygger på igenkänning av ords stavning, utan närmar sig språket matematiskt, kan den arbeta med vilket språk som helst.

DRE (Dynamic Reasoning Engine) är kärnan i Autonomymjukvaran. Det är ett slags databas där ett "avtryck" av varje dokument sparas. När Autonomy väl har identifierat nyckelkoncepten i ett dokument kodas termernas (de termer som ingår i det aktuella dokumentet) underliggande mönster in i mjukvaran. Dessa "mjukvaruabstracts" kallas för "Concept Agents". De används sedan för att känna igen mönstren för liknande koncept och hitta träffar i vilken annan text som helst. "Concept Agents" skapas genom att DRE:n får köra igenom en textmängd. Texten kan antingen vara en rad som skrivs in manuellt med specifik information för att träna agenten, ett existerande dokument eller en hel dokumentsamling. Agenten tittar på källtexten och beräknar mönstren för de viktigaste koncepten i texten. Man kan finjustera agenten genom att låta den "läsa" ytterligare text. För varje ny text "rättar" agenten till sitt fokus. Denna "re-training" är viktig – den höjer träffsäkerheten hos agenten.

Istället för att söka efter exakta nyckelordsmatchningar kan "Concept Agents" söka efter liknande idéer. Idén, eller konceptet, representeras av de termmönster och kontextuella förhållanden som vanligtvis är viktiga i de dokument som innehåller konceptet i fråga. Mönstren är inte beroende av språkliga strukturer eller semantik och påverkas därför inte av slang eller grammatiska eller regionala variationer i språket. Agenter är effektivast när de är väldigt specialiserade på ett specifikt koncept.

Autonomy behandlar ord som abstrakta betydelseenheter och får fram deras betydelse genom ordens placering i sin kontext. När man för första gången börjar använda DRE:n har den redan en statistisk förståelse för mönstren i "normal" engelska, men maskinen kan tränas för att känna igen mönstren i vilket språk som helst.

"Concept Agents" kan användas för att automatiskt sortera in dokument i fördefinierade kategorier. Kategorierna kan definieras genom att träna en mängd "Concept Agents". Träningen består i att förse "Concept Agents" med exempel på dokument för varje kategori. Genom att träna systemet med representativa data som redan har kategoriserats manuellt lär sig systemet att känna igen vilka mönster som är statistiskt sett mest signifikanta i dokument som representerar en viss kategori. När man sedan låter systemet "läsa" en ny text kan det snabbt bestämma hur bra texten passar in på kategorierna. Detta kan göras antingen i realtid eller som en batchprocess.

Det standardpaket som används i varje Autonomytillämpning är The Knowledge Server, som består av Query Engine, Indexer samt User Interface. Dessutom kan man välja till vissa valfria moduler såsom Categorizer, Web Spider, Distributed Search och Knowledge Visualizer. För indexeringsverksamheten vid Riksdagsbiblioteket är modulen Categorizer av störst intresse.

www.autonomy.com

KTH, NADA

På Institutionen för numerisk analys och datalogi (NADA) vid Kungliga Tekniska Högskolan i Stockholm forskas det i språkteknologi. Hercules Dalianis och några av hans kollegor besökte Riksdagsbiblioteket i september 1999 och diskuterade våra önskemål kring automatisk indexering och eventuella lösningar som de kunde bistå med. Efter detta har Hercules Dalianis och en doktorand, Johan Carlberger, skapat ett demoprogram som vi har haft möjlighet att testa.

Nedan följer deras förklaringar till hur programmet fungerar och vilka möjligheter till förbättringar som finns:

”Vi har utvecklat ett program som med statistiska metoder extraherar ett godtyckligt antal indexord från riksdagens texter. För varje dokument väljs ett antal indexord utifrån följande kriterier:

- * Ordet finns i tesaurusen.
- * Hur ofta förekommande ordet eller någon av dess böjningsformer är i dokumentet relativt andra ord i dokumentet.
- * Hur ofta förekommande ordet eller någon av dess böjningsformer är i dokumentet relativt samma ord i hela textsamlingen.
- * Hur stor andel av samtliga dokument som innehåller ordet eller någon av dess böjningsformer.

Som en förekomst av ett tesaurusord räknas också ord som betecknas ”Use For” UF i tesaurusen.

För varje tänkbart index-ord beräknas ett mått på hur ”bra” ordet skulle vara som indexord genom att de olika kriterierna sammanvägs. Som utdata presenteras de bästa orden, antingen alla eller så många man önskar.

Dessa kriterier kan naturligtvis ändras och kompletteras under vidareutvecklingen av programmet. Givet en stor uppsättning texter med redan utvalda indexord kan vi optimera sammanvägningen av urvalskriterierna och därmed träna programmet till att välja indexorden så bra som möjligt.

Programmet använder effektiva algoritmer för extrahering av statistik och ordböjning och kan analysera texter med en hastighet av 1000-tals ord per sekund.

Förslag på möjliga förbättringar av programmet:

- * Möjlighet att välja indexord bland egennamn och eventuellt icke-tesaurusord.
- * Extrahering av flerordsuttryck ur tesaurusen. Just nu används endast enkla tesaurusord.
- * Analysering av sammansatta ord för att hitta tesaurusord som förekommer som sammansättningsled.
- * Full användning av informationen i tesaurusen, t.ex. BT, NT.
- * Analysering av tidigare indexerade dokument för att kunna välja tesaurusord som inte finns explicit i ett dokument.
- * Användning av dokumentens typografiska och strukturella utformning för att få ytterligare urvalskriterier.”

www.nada.kth.se

4.3 Sammanfattning – marknaden

Automatisk indexering, och dess övergripande område informationssökning, är idag inne i ett mycket expansivt skede. I och med den explosionsartade ökningen av antalet elektroniska dokument på senare år, både på Internet men också i företags och organisationers intranät, har behovet av att på ett effektivt sätt kunna söka och återfinna dokument ökat och därmed har forskningen inom informationssökning intensifierats. Det leder i sin tur till att marknaden också snabbt utvecklas med nya företag och produkter som dyker upp med allt kortare intervall.

Som de tidigare sidorna har visat finns det många företag, men det finns få vars produkter verkligen klarar av svenska och som är producerade just för indexeringsverksamhet. Eftersom området är så dynamiskt och expansivt är det dock rimligt att tänka sig att detta förhållande kommer att förändras. Sannolikt kommer marknaden att se annorlunda ut inom några år. Vissa företag har kanske försvunnit, andra har tillkommit och forskningen och produktutvecklingen har kommit längre. Förhoppningsvis kommer det då också att finnas fler och bättre produkter för sådana indexeringsbehov som Riksdagsbiblioteket har.

5 Urval

De fyra program som ingår i de tester som gjorts är producerade av tre språkteknologiska företag samt en forskningsinstitution. För enkelhetens skull benämns alla fyra *företag*, även om det inte är en rättvisande benämning för dem alla.

De företag som ingår i de tester som gjorts är följande:

Lingsoft
Conexor
LexWare Labs
NADA, KTH

Lingsoft, Conexor, LexWare Labs och NADA, KTH valdes ut för testning efter beaktande av följande kriterier:

Hantering av svenska

Som visats ovan finns det ett antal produkter som skulle kunna vara intressanta för indexeringsändamål. Som också framgått finns det inte många produkter som i dagsläget säger sig hantera svenska. Möjligheten att översätta och lokalisera befintliga produkter finns ju givetvis, men eftersom den testning som gjorts inom den här delstudien skulle utföras inom loppet av elva månader och till rimliga kostnader ansågs det nödvändigt att de produkter som skulle testas redan hanterade svenska.

Tidsaspekt

Den tid som fanns för att sondera marknaden och testa produkter var alltså elva månader. De produkter som skulle testas fick således inte ta för mycket tid att ”starta upp”, utan skulle helst gå att köra utan någon inlärningsperiod, vare sig för användaren eller för produkten. Program som t.ex. IBM:s Intelligent Miner for Text och Data Harmonys Machine Aided Indexer skulle ha tagit lång tid att anpassa till riksdagens material för att få tillförlitliga resultat. Dessutom skulle dessa produkter ha krävt en lokalisering till svenska och det betraktades som ett alldeles för stort och tidskrävande arbete för att passa för ett uppdrag inom projektet Automatisk indexering.

Geografisk aspekt

De företag som slutligen ingick i testningen ligger alla relativt nära, geografiskt sett. Conexor och Lingsoft återfinns i Helsingfors, Finland, LexWare Labs i Göteborg och NADA, KTH finns i Stockholm. Att anlita företag som ligger långt bort, t.ex. i USA, och som kanske inte ens har återförsäljare i Europa, bedömdes som osäkert.

Plattform

Om produkterna skulle gå att testa i riksdagen var det viktigt att de fungerade på den plattform som används i riksdagen, nämligen NT/Windows. Tyvärr installerades inget av de fyra företagens program på riksdagen. Av olika anledningar gjordes testkörningarna av KTH, Lingsoft och LexWares produkter hos företagen själva. Conexors produkt testades via Internet. Produkternas plattformstillhörighet visade sig således inte vara av någon betydelse.

Företagens vilja till anpassningar

Alla de fyra företag vars produkter slutligen testats har gjort någon slags anpassning av sina tekniker och instrument för att tillgodose indexeringsverksamhetens önskemål och behov. Eftersom inget färdigt program fanns att tillgå värderades företagens vilja att göra anpassningar högt.

Programmets tillämpningsområde

Då delstudien var begränsad både i tid och omfång var det viktigt att hitta program som indexerade och inget annat. Autonomys produkt var mycket intressant och vi skulle gärna sett att den hade blivit föremål för testning. Den är dock en mycket omfattande produkt, där den del som skulle kunna användas för indexering endast utgör en liten modul, och delvis på grund av detta ansågs den inte passa inom projektets ramar⁸

Sammanfattning - urval

Kriterierna som presenterats ovan var de som i huvudsak beaktades vid urvalet av produkter. Det fanns, som nämnts, andra produkter som var av intresse, men de föll bort på grund av en eller flera orsaker. Sammanfattningsvis kan sägas att det största hindret för att hitta intressanta produkter är att Riksdagsbibliotekets indexeringsverksamhet är begränsad till produkter som hanterar svenska. Om riksdagsstrycket hade varit skrivet på engelska så skulle marknaden definitivt ha varit mycket större. Man kan konstatera att svenska är ett litet språk och att de produkter som trots allt finns för svenska oftast befinner sig på ett utvecklingsstadium fortfarande.

⁸ En testning av Autonomy hade kunnat komma i intressekonflikt med det på riksdagen pågående URIS-projektet, som vid tidpunkten för den här delstudien var inne i en känslig fas.

6 Utvärdering av indexeringsprogram

Vid en jämförande studie mellan en manuell arbetsuppgift och dess automatisering kan många aspekter komma i fråga – ekonomiska, tidsbesparande, arbetsmiljömässiga, etc. Den här delstudien inom projektet Automatisk indexering koncentrerades dock på den språkliga aspekten och jämförde de termer som indexerare tilldelar dokument och de termer som olika datorprogram tilldelar desamma.

6.1 Metod

Det naturliga sättet att utvärdera en maskinell process som också görs manuellt är att jämföra den maskinella och manuella processens resultat, i detta fall de termer som processen genererar. Detta val gjordes även för den här delstudien. Underlaget för jämförelsen bestod av ett urval av riksdagstrycket, se nedan 6.2.

Vid både manuell, datorstödd och automatisk indexering tilldelas ett antal termer till ett dokument. Med termer avses fortsättningsvis ord och fraser som används för att beskriva ett dokumentets innehåll. Term används som övergripande och generellt begrepp för följande grupper:

- De ord och fraser som återfinns i Riksdagens tesaurus. Dessa utgör ett kontrollerat språk och kallas i rapporten fortsättningsvis *tesaurusdeskriptorer*.
- De ord och fraser som anges som icke-deskriptorer i Riksdagens tesaurus. I denna rapport kallas de fortsättningsvis *icke-deskriptorer*.
- De ord och fraser som anger namn på geografiska platser, företag samt statliga verk och myndigheter. Dessa kallas fortsättningsvis *namnord* i rapporten.
- De ord och fraser som inte ingår i någon av de ovan nämnda grupperna, men som ändå är innehållsrika ord och fraser och som kan användas för innehållsbeskrivning av dokument. I föreliggande rapport kallas de fortsättningsvis *innehållsord*.

Endast i de fall då specificering av de olika grupperna är nödvändig används de olika benämningarna. I övrigt används benämningen *term*.

För att kunna utvärdera de termer som indexeringsprogrammen gav till dokumenten i testmaterialet togs ett facit fram genom att två indexerare vid sektionen för indexering och registerproduktion också fick indexera testmaterialet (se nedan, 6.2 och 6.3). Programmen utvärderades med avseende på antalet överensstämmelser med en indexerare i valet av termer. Detta sätt att utvärdera indexeringprogram har använts i flera liknande tidigare studier (Turney 1997, 1999 och Bureau van Dijk 1995).

Med hjälp av en matris kan man lätt få överblick över hur termer kan bedömas vid en jämförelse mellan indexerare och indexeringsprogram.

	Bedömd som relevant term av en indexerare	Bedömd som icke-relevant term av en indexerare
Bedömd som relevant term av ett program	A	B
Bedömd som icke-relevant term av ett program	C	D

I fält A hamnar de termer som bedöms som relevanta av både indexeraren och programmet och i D de som bedöms som icke-relevanta av dem båda. I de fall indexeraren och programmet inte har gjort samma bedömning hamnar termen i antingen fält B eller C. Ju fler termer som hamnar i fält A och D, desto bättre överensstämmelse mellan programmet och indexeraren och ju fler termer som hamnar i fält B och C, desto sämre överensstämmelse mellan programmet och indexeraren.

För att jämföra indexerarnas och programmets förslag på termer användes ett allmänt tillgängligt program som finns på <http://instruct.uwo.ca/gplis/677/indecons.html>. Det gör en mekanisk jämförelse mellan två mängder av termer. Programmet beräknar *indexeringskonsistensen* (andelen överensstämmande termer) mellan de två termmängderna. Dessutom anges antalet termer för varje mängd, det totala antalet termer för båda mängderna tillsammans samt hur många och vilka termer som förekommer i båda mängderna. Dessa uppgifter användes för att räkna ut ytterligare utvärderingsparametrar med vars hjälp man kan bedöma ett indexeringsprogramns kvalitet. De utvärderingsparametrar som räknades fram var *recall*, *precision* och *F-värde* (van Rijsbergen 1979). Recall beskriver täckning, precision träffsäkerhet och F-värdet anger ett sammanvägt värde av de andra två parametrarna och kan således underlätta vid utvärderingen av de två andra parametrarnas värden. När värdena för precision och recall ligger på ungefär samma nivå blir F-värdet i princip samma sak som medelvärdet av precision och recall. När precisions- och recallvärdena är olika blir F-värdet lägre än medelvärdet (Turney 1999:9). Parametrarna beräknas med följande formler:

$$\text{Indexeringskonsistens: } \frac{|A \cap B|}{|A \cup B|}$$

$$\text{Recall: } \frac{|A \cap B|}{|B|}$$

$$\text{Precision: } \frac{|A \cap B|}{|A|}$$

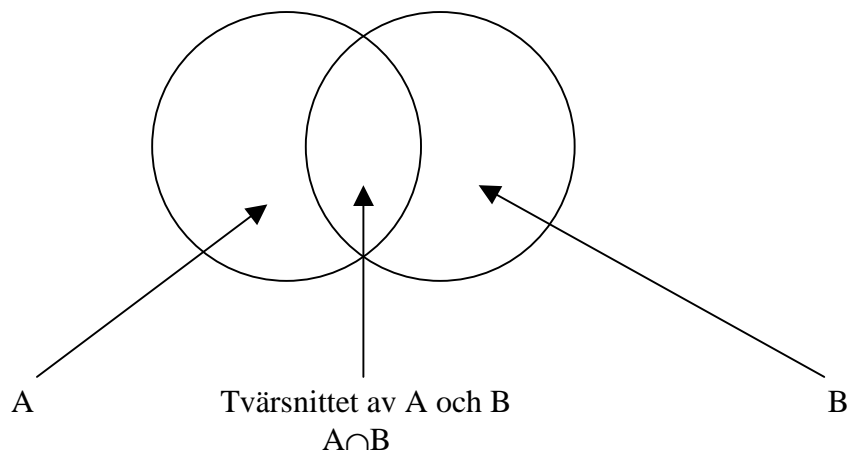
där:

A: programmets förslag på termer

B: relevanta termer enligt indexerarna

\cup : unionen av A:s och B:s termer, d.v.s. den sammanlagda mängden av de termer som finns i någon av de båda mängderna

\cap : tvärsnittet av A:s och B:s termer, d.v.s. de termer som finns i både A och B (de överensstämmande termerna).



$$\text{F-värde: } \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

6.2 Material

Materialet som användes för jämförelsen är ett urval av de 4699 dokument som indexerades för riksmötet 1997/98. Testmaterialet bestämdes till att utgöra 4 % av denna dokumentmängd. Sålunda ingick 187 dokument i undersökningen. De dokumenttyper som indexerades vid sektionen för indexering och registerproduktion är, som nämnts i kapitel 2.2, propositioner, skrivelser, förslag, redogörelser, motioner, interpellationer och frågor.

Antalet dokument fördelade sig enligt följande för de olika dokumenttyperna:

Dokument-Typ	Totalt antal dokument	Avrundat antal dokument som representerar dokumenttypen i undersökningen
Propositioner	153	6
Skrivelser	37	1
Interpellationer	313	13
Frågor	903	36
Förslag	12	0
Redogörelser	10	0
Motioner, allmänna	2495	100
Motioner, följd	776	31

Riksdagens utredningstjänst var behjälpliga och bedömde de siffror som räknades ut för respektive dokumenttyp.

Testmaterialet skulle i största möjliga mån avspegla den verkliga dokumentssituationen, med de rätta proportionerna mellan de olika dokumenttyperna. För att få fram rättvisande proportioner valde vi att utgå från dokument som producerats under ett enda riksmöte. Riksmötet 1997/98 ansågs som representativt efter att ha jämförts med riksmötena 1995/96 och 1996/97.

Ett annat önskemål var att så många dokumenttyper som möjligt skulle ingå i testmaterialet. Vi plockade ut samma andel dokument ur varje dokumenttyp (4 %) och det minsta antal dokument per dokumenttyp som kunde plockas ut var ett dokument. Då antalet dokument är mycket ojämnt fördelat mellan de olika dokumenttyperna resulterade detta urvalsförfarande i att vissa dokumenttyper inte representeras i testmaterialet. 4 % av t.ex. det totala antalet redogörelser (10) blir ju bara 0,4 dokument. De dokumenttyper som således bestod av för få dokument är förslag och redogörelser. De liknar dock propositioner till sin struktur, utformning och språk och därför ansågs dessa dokumenttyper kunna uteslutas ur testmaterialet utan alltför stor förlust.

Från början fanns en tanke att utforma testmaterialet efter en tidsmässig bedömning. Testkorpusen skulle helt enkelt utgöras av det material som en indexerare hinner indexera på en vecka (eller två). Men det fanns ingen tillgänglig statistik över hur många dokument som en indexerare hinner med på denna tid och det visade sig vara mycket svårt att uppskatta en sådan siffra, då det kan variera mycket beroende på flera faktorer (dokumentets typ, längd, svårighetsgrad, indexerarens vana och skicklighet etc.). Utplockning av material enligt denna metod skulle troligtvis inte heller ha avspeglat proportionerna mellan dokumenttyperna särskilt väl. Detta förslag fick således stå tillbaka för det mer abstrakta och arbiträra sättet att välja en lämplig siffra som skulle ge ett tillförlitligt undersökningsmaterial.

Utplockningen av materialet gjordes manuellt och relativt slumpvis. I stort sett togs vart 25:e dokument ut ur varje dokumenttyp från riksmötet 1997/98 för att tillhöra testkorpusen. Viss

hänsyn togs dock till dokumentens partitillhörighet och längd. Detta gjordes för att undvika att ett visst parti, och därmed dess politiska språk, skulle bli överrepresenterat i testmaterialet och för att spegla dokumentens längdskillnader – vissa motioner är en mening långa, andra är tiotals sidor. Beträffande motionerna togs hänsyn även till vilket utskott som hade behandlat dem och till deras ämnesområde.

6.3 Den manuella indexeringen

Den ovan beskrivna urvalsprocessen resulterade i listor med dokumentnamn och -nummer. Dessa listor gavs till två av de fyra indexerarna på sektionen (indexerare X och Y) och deras arbete bestod sedan i att indexera detta testmaterial med tesaurusdeskriptorer och namnord (se ovan 6.1). Under tiden denna dubbelindexering pågick fick de två indexerarna inte diskutera dokumenten eller indexeringsfrågor som rörde dessa med varandra; detta för att resultaten från dubbelindexeringen skulle bli så rättvisande som möjligt. Indexerarna skulle indexera oberoende av varandra och inte bli influerade av varandras indexering. De hade inte heller tillåtelse att söka i sakregistret (se kapitel 2.2) för att se hur de i testkorpussen ingående dokumenten indexerats. Indexerarna tilldelade således, var och en för sig, ett antal tesaurusdeskriptorer och namnord till vart och ett av dokumenten i testkorpussen. För varje dokument fanns det till slut således två listor med termer – en för varje indexerare.

De termlistor som blev resultatet av de två indexerarnas arbete jämfördes med varandra med hjälp av programmet som beskrevs under 6.1. Syftet med denna jämförelse var att ge en uppfattning om hur enhetlig den manuella indexeringen är och att ge ett facit till jämförelserna med datorprogrammen. Antagandet som jämförelsen baseras på är att ingen av indexerarnas termer är bättre (speglar bättre ett dokumentets innehåll) än den andras, men att den mänskliga indexeringen totalt sett är det facit som programmens utdata bör jämföras med; ju mer likt den mänskliga indexeringen, desto bättre.

Indexeringskonsistensen mellan de två indexerarna är 34 %. Den mängd termer som användes totalt av indexerarna var heterogen till sin karaktär och bestod av många termer. Ibland överensstämde inte någon term, utan indexerarna hade valt helt olika termer för ett och samma dokument. I dessa fall blev den totala term mängden indexerare X:s termer + indexerare Y:s termer och detta antal kunde vida överskrida det som på sektionen för indexering och registerproduktion används som riktmärke för de olika dokumenttyperna - propositioner bör ej indexeras med mer än 20 termer, motioner med inte fler än 10 och frågor och interpellationer bör inte indexeras med fler än 5 termer. Eftersom den totala term mängden för varje dokument var så heterogen och ojämn kan den knappast sägas vara ett bra facit för en jämförelse. Därför har programmens termlistor inte bara jämförts med indexerarnas totala antal termer, utan även med var och en av indexerarnas term mängder.

En indexeringskonsistens på 34 % mellan två indexerare är inte ett högt värde. En förklaring till resultatet är att de två indexerare som gjorde dubbelindexeringen är relativt nya – indexerare X hade vid tillfället för jämförelsen arbetat med indexering i ca ett år och indexerare Y bara ca tre månader. Enligt den utvärdering av automatisk indexering som gjordes vid Europaparlamentet 1995 kan man mellan två nybörjarindexerare förvänta sig en indexeringskonsistens på mellan 20 och 30 %. För vana och skickliga indexerare kan dock

indexeringskonsistensen uppgå till mellan 60 och 80 % (Bureau van Dijk 1995:78). Mot bakgrund av dessa uppgifter är 34 % inte ett påfallande lågt värde.

Ytterligare en orsak till den relativt låga siffran 34 % ligger i användandet av jämförelseprogrammet (se ovan 6.1). Programmet jämför sträng mot sträng och tar ingen hänsyn till att vissa av orden i de två termmängderna trots grafisk olikhet kanske faktiskt tillhör samma semantiska hierarki i tesaurusen. Tesaurusen plattas således ut och dess sinnrika hierarkiska konstruktion som anger relationer åt olika håll kommer inte till sin rätt. Inte sällan använder sig en indexerare av en BT, medan den andra använder sig av en NT till den förres BT. I motion 9798Ju217 använder t.ex. indexerare X deskriptorn ”Kommittéer” som är en BT till deskriptorn ”Parlamentariska utredningar” vilken är den deskriptor som indexerare Y använder för samma motion. Genom att beakta detta fenomen skulle likheten mellan de två indexerarnas termmängder säkerligen blivit större, men en dylik jämförelse hade blivit avsevärt mer komplicerad och tidskrävande än vad studien medgav.

6.4 Förutsättningar för testerna

Alla fyra programmen testades med samma material – testmaterialet bestående av 187 dokument. Innan dessa dokument testades fick dock företagen tillgång till visst material från riksdagen för att kunna göra vissa anpassningar och för att i viss mån ”träna” sina program. Alla utom Conexor har fått tillgång till 144 dokument för att få en uppfattning om hur riksdagsstrycket ser ut. LexWare Labs och KTH har dessutom haft tillgång till Riksdagens tesaurus. Lingsoft fick, vid sin företagspresentation på Riksdagsbiblioteket, bekanta sig med tesaurusen. Conexor har varken haft tillgång till tesaurusen eller till det omfattande provmaterialet, men i gengäld har de vid två tillfällen testkört två dokument ur riksdagsstrycket och sedan visat upp resultatet av sitt program. Detta resultat har vi tittat på, kommenterat och sedan skickat tillbaka till Conexor. På så sätt har de fått grundlig information om vad Riksdagsbiblioteket önskar sig av en automatisk indexering.

Alla kandidater har alltså vetat om att den manuella indexeringen använder sig av Riksdagens tesaurus i sitt arbete. Något som dock inte tydligt framgått i kommunikationen med kandidaterna är att även namnord används i indexeringsarbetet vid sektionen för indexering och registerproduktion.

Från början var meningen att all testning skulle utföras på riksdagen. Av olika skäl blev detta inte fallet. Både LexWare Labs, KTH:s och Lingsofts program kördes i stället på respektive företag och resultaten skickades sedan till Riksdagsbiblioteket för utvärdering. Detta sätt att testa har inneburit att projektet Automatisk indexering inte har haft full kontroll över dessa tester och att resultaten av testerna måste ses mot bakgrund av detta. Conexors produkt installerades inte heller på riksdagen, utan testades via Internet. Conexor lade upp sitt program på sidan <http://condemo.co.helsinki.fi/proj/riksdagen/> och projektet Automatisk indexering fick tillgång till programmet med hjälp av användarnamn och lösenord.

I följande avsnitt redovisas varje kandidats resultat för dokumenttypen följdmotioner. Detta är en vanlig dokumentgrupp som till sin karaktär ofta är en blandning av allmänna motioner och

propositioner. De är motioner, d.v.s. skrivna av enskilda eller flera ledamöter, men behandlar ämnen som presenterats i en proposition och således återfinns många av propositionens ord och mycket av dess ”manglade” språk i följdmotionerna. Resultaten från de övriga dokumenttyperna återfinns i bilaga Lingsoft, bilaga Conexor, bilaga LexWare Labs och bilaga KTH.

6.5 Resultat

6.5.1 Lingsoft

Lingsofts output är en lista av de substantiv och substantivfraser som finns i ett dokument. De är rankade efter förekomst och de termer som har högst frekvens i dokumentet presenteras överst i listan. Ju fler substantiv och substantivfraser ett dokument innehåller desto längre blir listan. Den innehåller både tesaurusdeskriptorer, icke-deskriptorer, namnord och innehållsord.

Eftersom Lingsofts utdata ofta består av väldigt många termer och eftersom många av dem inte är tesaurusdeskriptorer har den efterbearbetats. Varje dokuments output har beaktats ur tre aspekter:

- hela listan
- de tio högst rankade termerna
- de tesaurusdeskriptorer som finns i listan

Anledningen till att de tio högst rankade termerna är en aspekt för beaktande är att motioner, som nämnts under 2.2 och 6.3, inte bör indexeras med fler än tio termer.

För att kunna göra en rimligare jämförelse mellan programmen valde vi att filtrera Lingsofts termer genom Riksdagens tesaurus. Efter filtreringen bestod resultatlistan bara av tesaurusdeskriptorer. Filtreringen gör inte Lingsofts program full rättvisa då namnorden inte kommer med, men inom den här undersökningen var det den bästa lösningen för att kunna jämföra företagen på ett likvärdigt sätt.

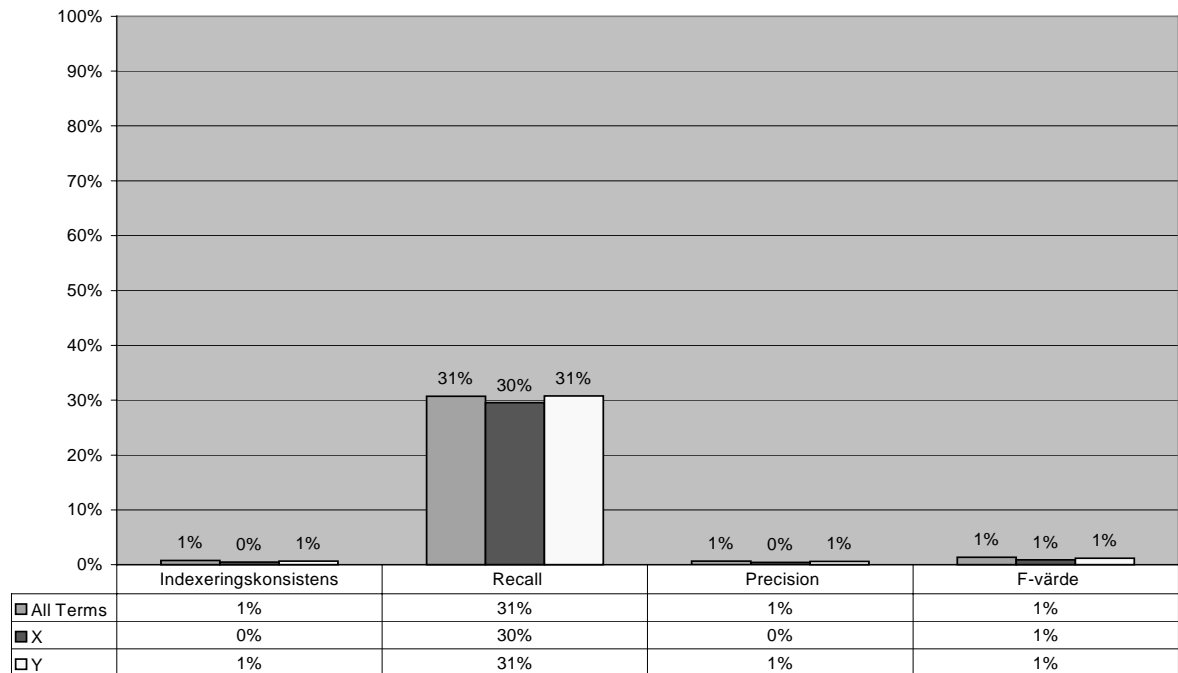


Diagram 1: Följdotioner, Lingsofts alla och indexerarnas termer

Det första diagrammet beskriver jämförelsen mellan hela resultatlistan och indexerarnas olika term mängder (indexerare X:s termer, indexerare Y:s termer och deras sammanlagda term-mängd (All Terms)). De siffror som anges är medelvärden för de 31 dokument som ingår i gruppen följdotioner. Här kan vi se hur indexeringskonsistens, precision och F-värde antar mycket låga värden. Recallvärdet når upp till 30-procentstrecket. En anledning till att indexeringskonsistens, precision och F-värde får så låga värden är att Lingsofts resultatlista består av så många termer. Även om programmet hittar nästan en tredjedel (recall ca 30 %) av de termer som indexerarna har bedömt som relevanta för dokumenten så blir värdena för indexeringskonsistens och precision ändå obefintliga, eftersom programmet plockar fram mängder med icke-relevanta termer.

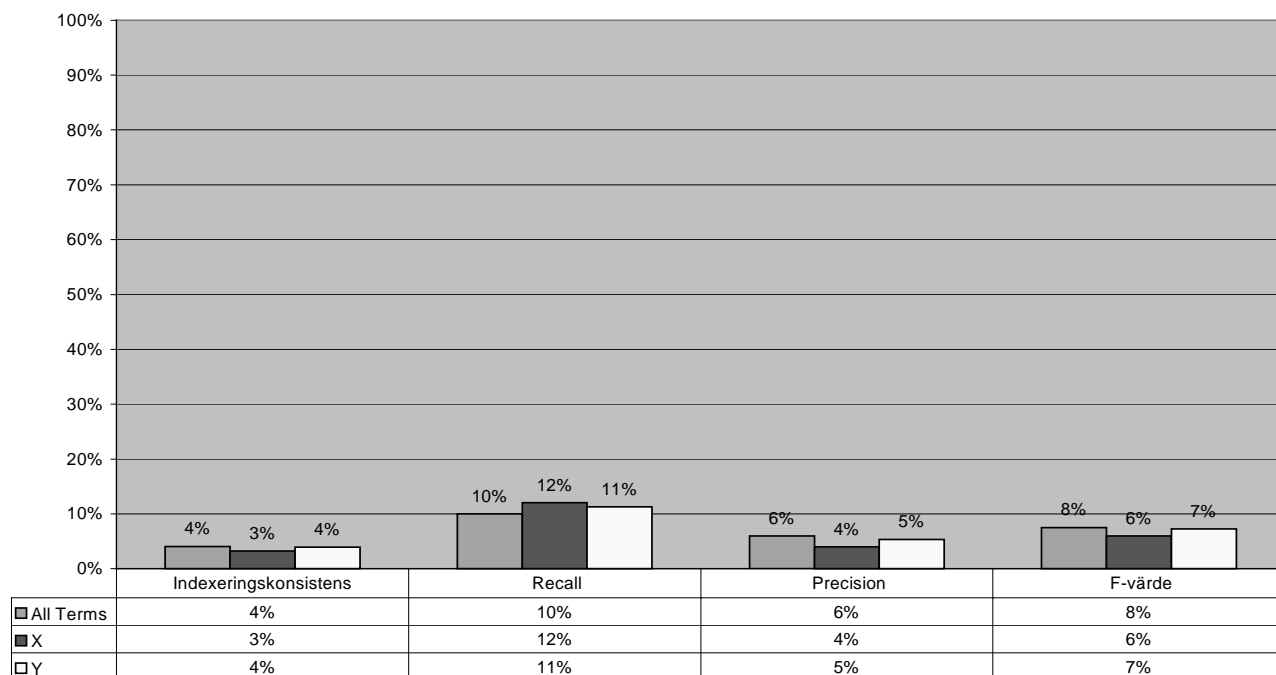


Diagram 2: Följdnotioner, Lingsofts tio högst rankade och indexerarnas termer

I det andra diagrammet visas de resultat som fås efter att bara ha beaktat de tio högst rankade termerna i Lingsofts resultatlista. I detta diagram uppvisas ett jämnare resultat parametrarna emellan. Recallsiffrorna innebär att programmet hittar ungefär en tiondel av de termer som indexerarna bedömt som viktiga. Samtidigt säger värdena för precision att i genomsnitt är inte ens en term bland de tio överst rankade termerna i Lingsofts resultatlista relevant enligt indexerarna. Att recallvärdena är högre än precisionvärdena beror på att indexerarna i genomsnitt använder färre termer än tio för att indexera dokumenten. Precisionvärdena i den här jämförelsen bygger på att antalet termer som utgör tvärsnittet av indexerarnas termer (X, Y respektive All Terms) och de tio högst rankade termerna i Lingsofts resultatlista alltid delas med tio.

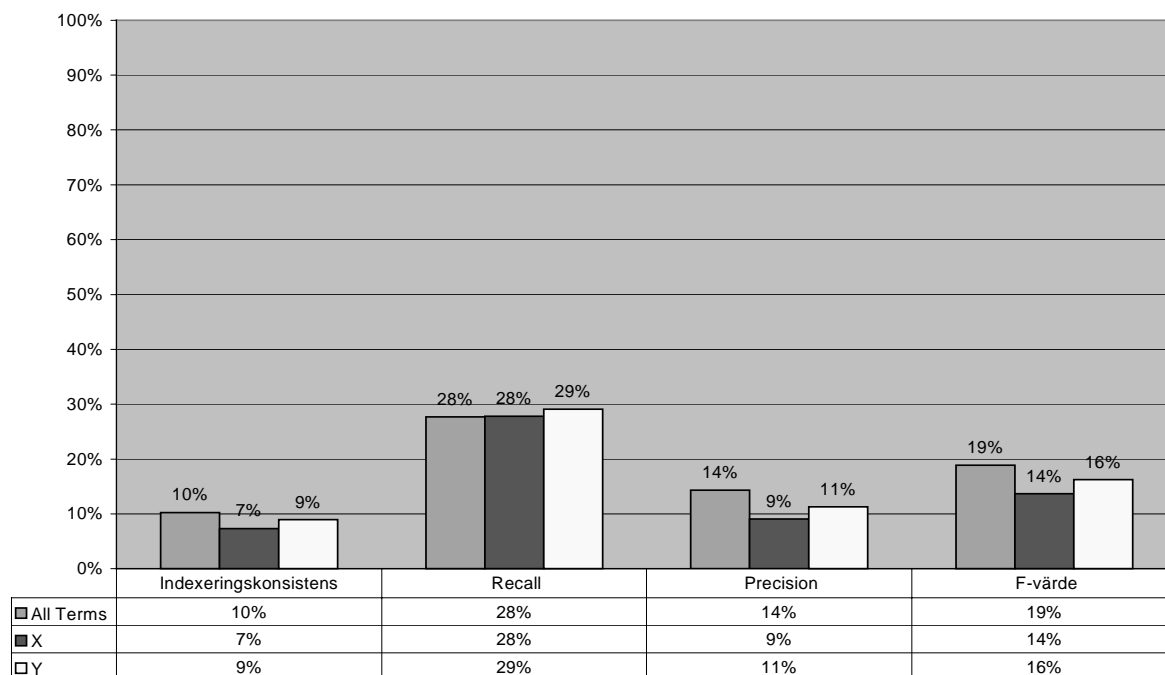


Diagram 3: Följdotioner, Lingsofts filtrerade och indexerarnas termer

I det tredje diagrammet visas de resultat som fås efter att ha filtrerat Lingsofts resultatlista genom tesaurusen och på så sätt fått fram de termer i resultatlistan som också finns i Riksdagens tesaurus. Här är det lätt att se att värdena för alla parametrar utom recall har förbättrats radikalt jämfört med när hela resultatlistan jämfördes med indexerarnas term mängder. Anledningen till att recallvärdet har försämrats är att de namnord som ingick i den ursprungliga resultatlistan har försvunnit efter filtreringen genom tesaurusen. Recall är trots nedgången den parameter som får bäst resultat. Att All Terms (indexerarnas sammanlagda term mängd) får bäst resultat i indexeringskonsistensen beror på att All Terms ofta innehåller totalt sett fler termer än de enskilda indexerarnas term mängder gör och att det således blir större överensstämmelse mellan All Terms och Lingsofts förslag till termer än vad det blir mellan de enskilda indexerarnas termer och Lingsofts förslag.

Lingsofts resultat visar på det faktum att rangordning av termer efter deras frekvens i ett dokument ger inte ett tillräckligt bra underlag för att bedöma deras relevans. Andra metoder, såsom Inverse Document Frequency och jämförelse mot en referenskorpus, måste också användas för att tillförlitligare kunna bedöma en terms relevans i ett dokument.

6.5.2 Conexor

Conexors output är, liksom Lingsofts, en lista av de substantiv och substantivfraser som finns i ett dokument. Ju fler substantiv och substantivfraser ett dokument innehåller desto längre blir listan. Den innehåller både tesaurusdeskriptorer, icke-deskriptorer, namnord och innehållsord. Termerna i listan är rankade efter relevans och presenteras i relevansordning med de mest relevanta överst. Hur relevansen beräknas är Conexors företagshemlighet.

Eftersom Conexors resultatlista ofta består av så många termer och eftersom många av dem inte är tesaurusdeskriptorer har den efterbearbetats. Varje dokumentets output har beaktats ur tre aspekter:

- hela listan
- de tio högst rankade termerna i listan
- de tesaurusdeskriptorer som finns i listan

Anledningen till att de tio högst rankade termerna är en aspekt för beaktande är att motioner, som nämnts ovan under 2.2 och 6.3, inte bör indexeras med fler än tio termer.

På samma sätt som i fallet Lingsoft filtrerades Conexors termer genom Riksdagens tesaurus. Efter filtreringen hade alla program en resultatlista som bara bestod av tesaurusdeskriptorer. Filtreringen gör inte Conexors program full rättvisa, men inom den här undersökningen var det den bästa lösningen för att kunna jämföra företagen på ett likvärdigt sätt.

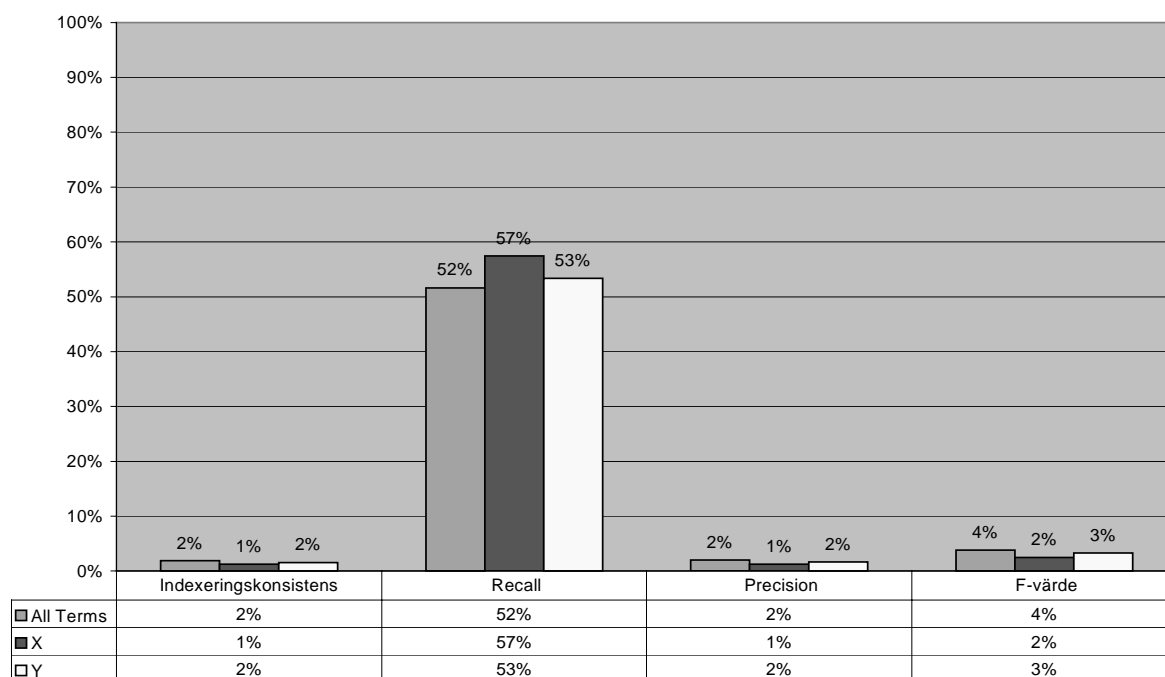


Diagram 4: Följdmotioner, Conexors alla och indexerarnas termer

Det första diagrammet visar resultaten av jämförelsen med Conexors hela resultatlista och indexerarnas termer. Liksom hos Lingsoft är resultaten för indexeringskonsistens, precision

och F-värde mycket låga och det beror också här på att resultatlistan ofta innehåller väldigt många termer. Överensstämmelsen mellan Conexors output och indexerarnas val av termer är helt enkelt inte särskilt stor. Conexors output är heller inte särskilt träffsäker – programmet hittar visserligen fler än hälften av de termer som indexerarna tilldelar dokumenten (recallvärdet), men det hittar också en mängd icke-relevanta termer, vilket gör att precisionen blir väldigt låg.

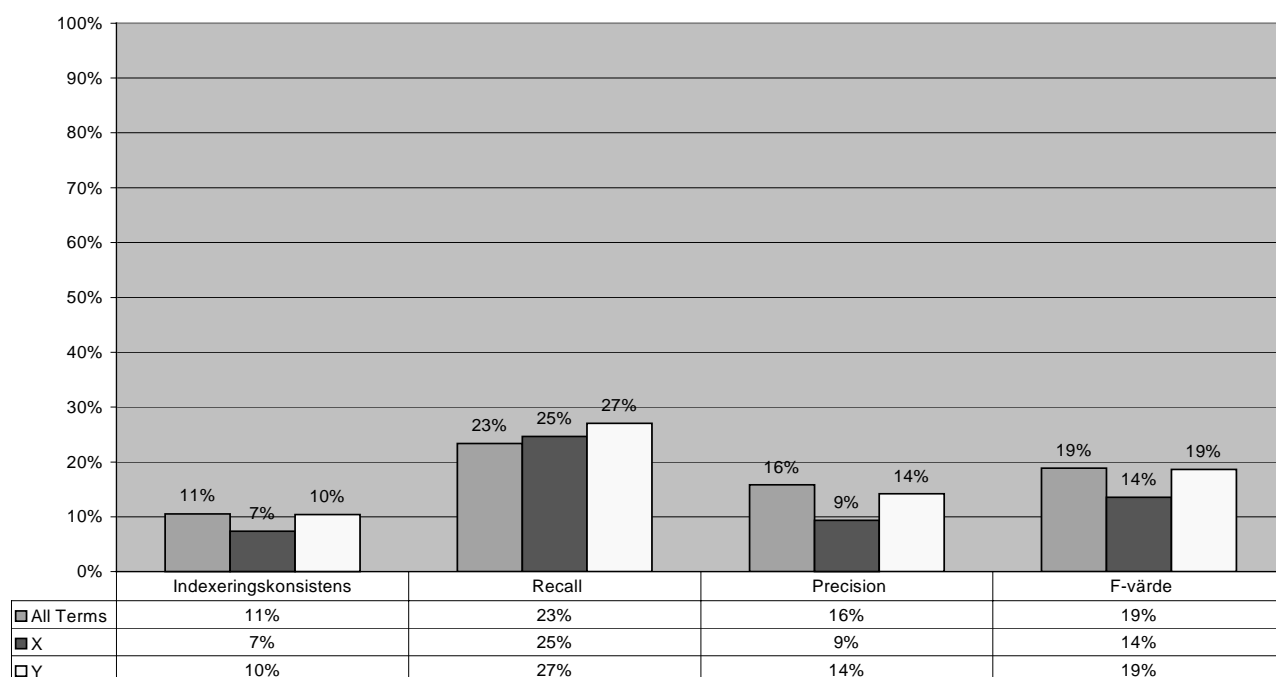


Diagram 5: Följdmotioner, Conexors tio högst rankade och indexerarnas termer

Det andra diagrammet visar de resultat man får då bara de tio högst rankade termerna i Conexors resultatlista beaktas. Värdena mellan de olika parametrarna blir mycket jämnare även om recall fortfarande får bäst resultat. Programmet hittar i genomsnitt en fjärdedel av de termer som indexerarna anser vara relevanta för en följdmotion (recallvärdet). Av de tio högst rankade termerna är i genomsnitt mellan 1 och 1,5 termer relevanta enligt indexerarna.

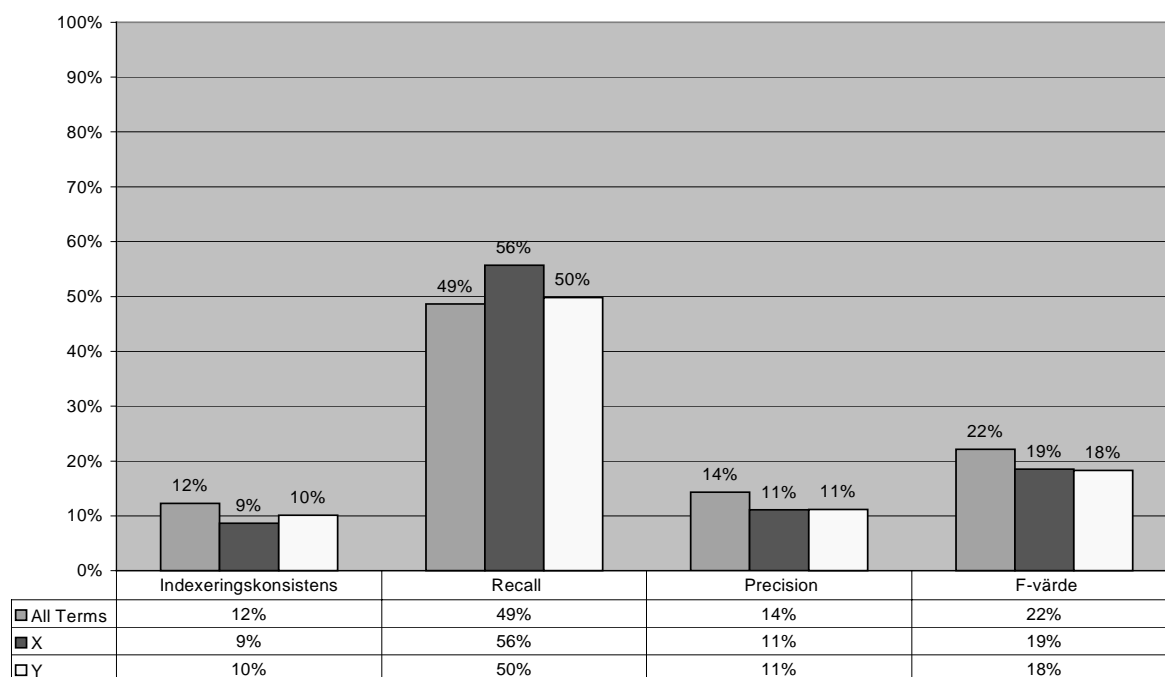


Diagram 6: Följd motioner, Conexor filtrerade och indexerarnas termer

Det tredje diagrammet visar de resultat man får efter att ha filtrerat Conexors output genom tesaurusen. Här jämförs således indexerarnas val av termer endast med tesaurusdeskriptorer ur Conexors output. Liksom hos Lingsoft förbättras värdena radikalt efter filtreringsoperationen jämfört med då hela resultatlistan jämfördes med indexerarnas term mängder, förutom för recall, vars värde sjunker något. Också här beror den försämringen på de namnord som faller bort vid filtreringen genom tesaurusen. Anledningen till att recallvärdet så gott som alltid blir bäst vid jämförelse med indexerare X beror på att indexerare X nästan alltid använder färre termer vid indexeringen än indexerare Y (och naturligtvis färre än den sammanlagda term mängden). Det är således lättare för programmet att hitta bara ett fåtal termer än att hitta fler korrekta termer, som jämförelsen med indexerare Y kräver.

6.5.3 LexWare Labs

LexWares Labs output består av en lista av tesaurusdeskriptorer. Ju fler ord som ingår i ett dokument, desto längre blir listan. Men listan innehåller inte alla de substantiv och substantivfraser som ingår i ett dokument som hos Lingsofts och Conexors output, utan programmet har bedömt vilka termer som är relevanta och sedan (i de fall de inte redan är tesaurusdeskriptorer) länkat dem till deskriptorer i tesaurusen. Resultatlistan består således bara av tesaurusdeskriptorer och detta har en klar inverkan på resultatet. Men även om resultatlistan oftast består av en mindre mängd tesaurusdeskriptorer så överskrids i allafall ibland de riktvärden för indexering av olika dokumenttyper som används på sektionen för indexering och registerproduktion. I de fall då programmet plockat fram fler än tio tesaurusdeskriptorer för en följd motion har således två jämförelser gjorts: en då hela resultatlistan beaktats och en där bara de tio högst placerade tesaurusdeskriptorerna beaktats.

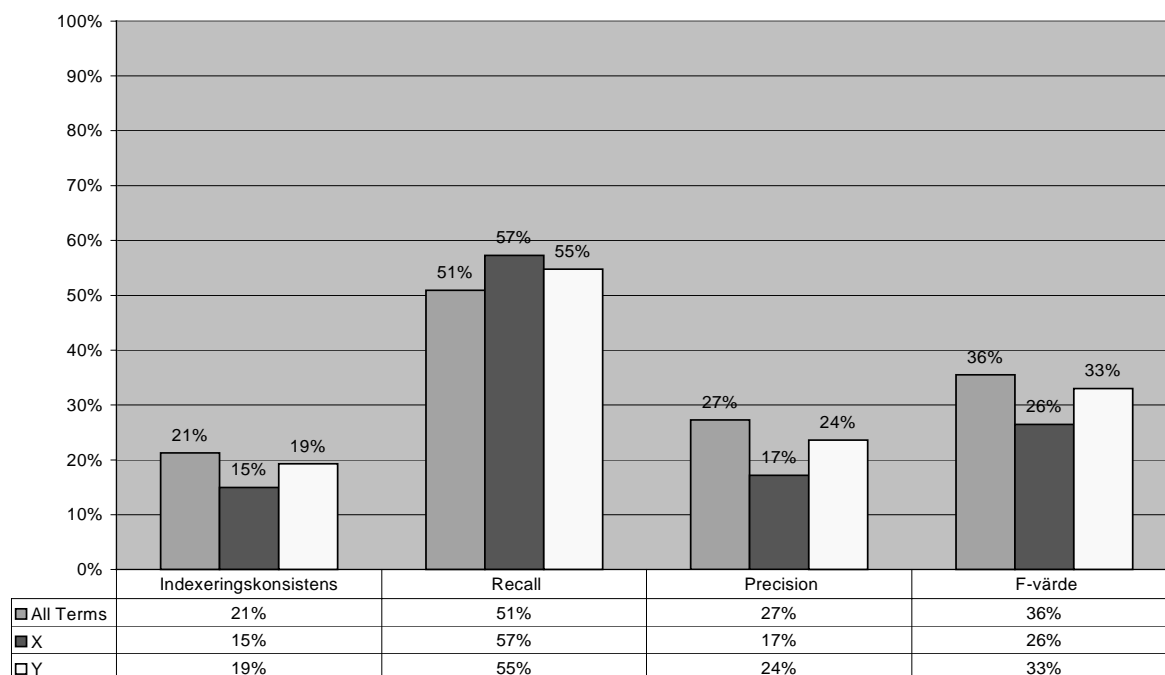


Diagram 7: Följdrotationer, LexWare Labs alla och indexerarnas termer

Det första diagrammet beskriver resultaten som erhålls när hela resultatlistan jämförs med indexerarnas termer. Recall får det högsta värdet, men det är inte så stora skillnader mellan de olika parametrarna som hos Lingsoft och Conexor.

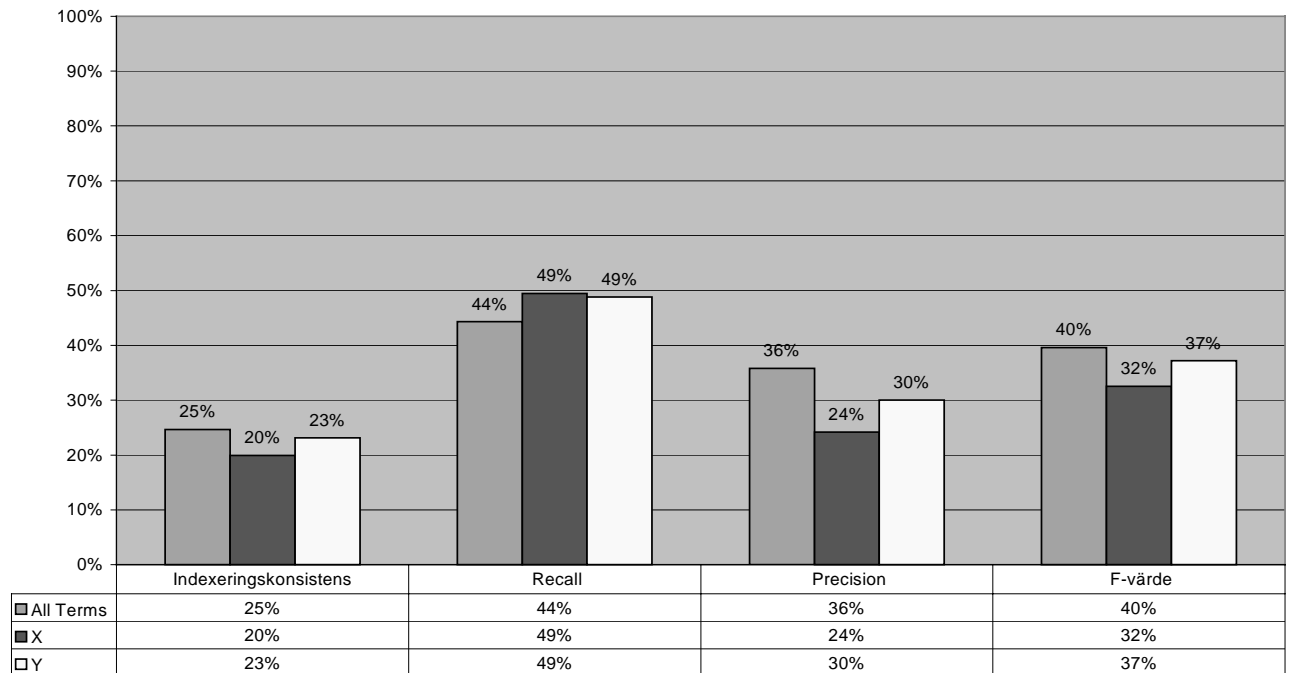


Diagram 8: Följdrotationer, LexWare Labs tio högst rankade och indexerarnas termer

Efter att man kapat resultatlistan och bara beaktat de tio högst placerade tesaurusdeskriptorer-na sjunker recallvärdet något, medan värdena för de andra parametrarna stiger något. I detta fall beror inte sänkningen av recallvärdet på att namnord har plockats bort ur LexWare Labs resultatlista, utan helt enkelt på att vissa av de termer som överensstämde mellan LexWare Labs förslag och indexerarnas val av termer fanns i resultatlistans nedre del, d.v.s. efter nummer tio i ordningen i listan.

6.5.4 KTH

KTH:s output består av fem förslag till termer per dokument. Förslagen består av tesaurusdeskriptorer och av de icke-deskriptorer som återfinns i tesaurusen. Förslagen presenteras i relevansordning. Hur relevansen beräknas beskrivs under 4.2. Eftersom KTH alltid ger fem termer i sin resultatlista har ingen efterbearbetning gjorts och KTH:s resultat kan således presenteras i ett enda diagram.

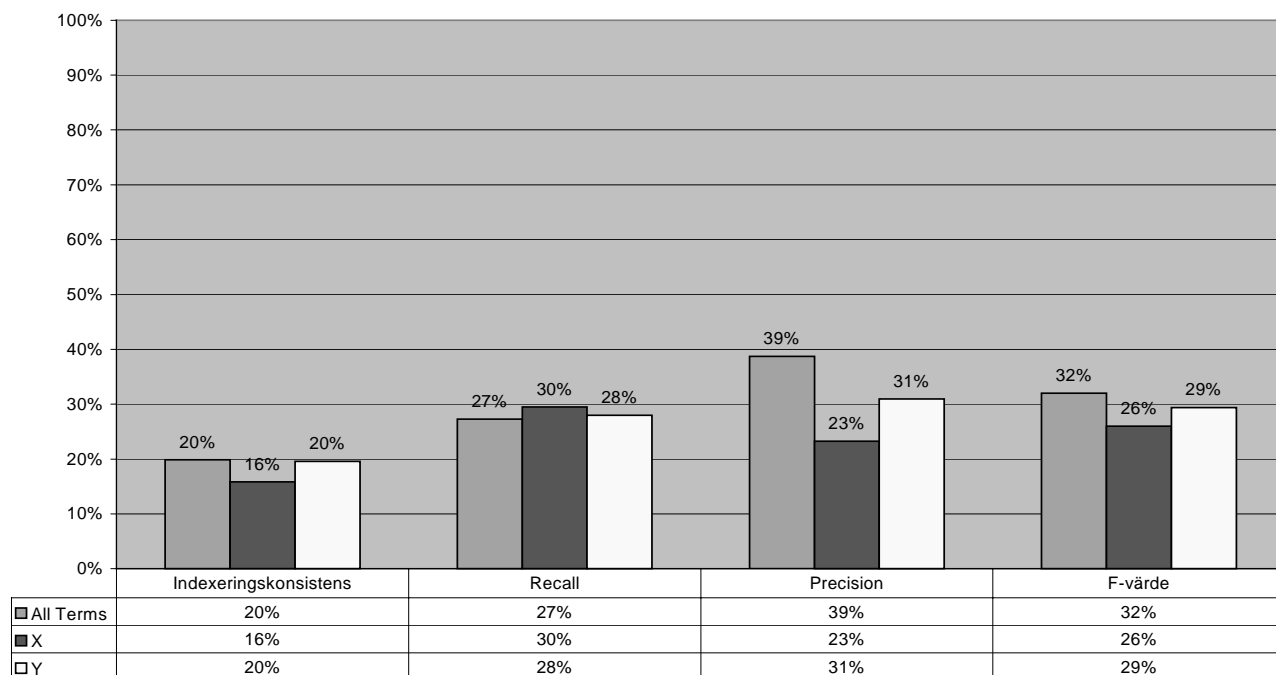


Diagram 9: Följdotioner, KTH:s alla och indexerarnas termer

Den parameter som får högst värde i jämförelsen mellan KTH:s resultatlista och indexerarnas termer är precision (utom i jämförelsen med indexerare X). Anledningen till att precisionen blir så hög är att KTH:s resultatlista bara innehåller fem termer. Även om t.ex. bara en term i KTH:s resultatlista också har föreslagits av indexerarna, och fyra av KTH:s förslag alltså inte överensstämmer med indexerarnas förslag så är det bara fyra termer som är felaktigt valda av KTH:s program. Detta är inte mycket när man t.ex. jämför med Conexors eller Lingsofts output, som kan bestå av hundratals felaktigt framplockade termer.

Som helhet kan sägas att de olika parametrarna får likartade värden. Detta syns bland annat i F-värdet, som i det här diagrammet verkligen ligger nära ett medelvärde av recall och precision. Indexeringskonsistens och precision blir lägre vid jämförelsen med indexerare X än vid jämförelsen med indexerare Y och All Terms. Recallvärdet är däremot högre vid jämförelse med indexerare X än vid jämförelsen med indexerare Y och All Terms. Detta beror på att indexerare X använder färre termer i sin indexering än vad indexerare Y och naturligtvis All Terms gör. Överensstämmelsen kan ju inte bli så stor om indexerare X t.ex. har använt två termer för ett dokument och KTH:s program bara bedömer en av dem som en relevant term och istället har valt ut fyra andra termer. Precisionen blir ju också låg, eftersom programmet genererar fler termer än vad indexerare X har använt. Att recallvärdet däremot blir ganska högt beror just på att indexerare X använder få termer i sin indexering. Om indexerare X bara har använt två termer och programmet hittar en av dem så blir det ett recallvärde på 50 % ($1/2=0,5=50\%$).

Det måste dock påpekas att många relevanta termer också går förlorade i och med att resultatlistan alltid består av fem termer. Den övre gräns som rekommenderas för indexering av motioner är tio termer per dokument. Även om indexerarna sällan tilldelar tio termer till ett dokument, så händer det ofta att fler än fem används.

6.6 Sammanfattning av testerna

Sammanfattningsvis kan sägas att kandidaternas resultat är av varierande kvalitet. Inget är dock så bra att det idag skulle gå att använda som ett fullgott alternativ till den manuella indexeringen. Däremot kan programmens utdata kanske tjäna andra syften än automatisk indexering.

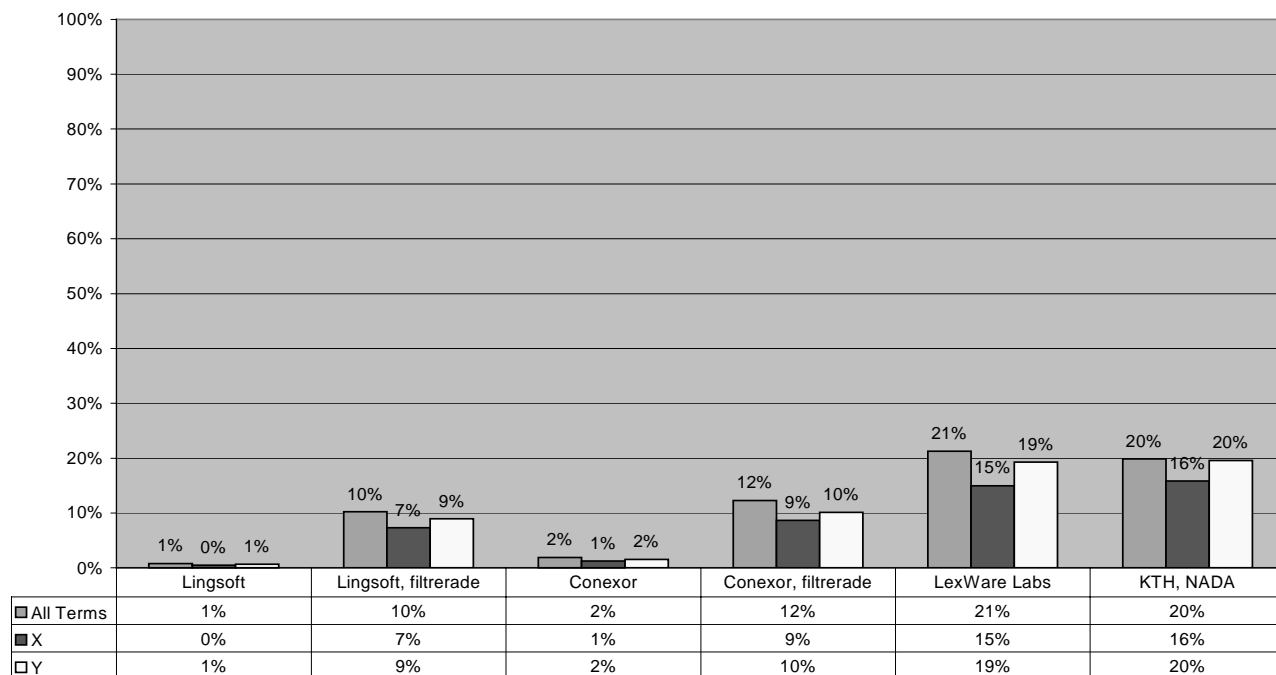


Diagram 10: Indexeringskonsistens, jämförelse mellan kandidaterna

De kandidater som uppvisar bäst resultat när det gäller indexeringskonsistens är LexWare Labs och KTH. Deras resultat ligger i genomsnitt strax under 20-procentstrecket och KTH:s är några tiondelars procentenheter bättre än LexWare Labs. Den undersökning som gjordes i Europaparlamentet 1995 (Bureau van Dijk 1995:73) fick snarlika resultat – de produkter som testades presterade mellan 16 % och 19 % i indexeringskonsistens. Man kan ju tycka att utvecklingen borde ha gått framåt på de drygt fyra år som gått sedan den undersökningen gjordes och att de resultat som framkommit i projektet Automatisk indexerings undersökning således borde vara bättre. Att så inte är fallet beror troligtvis på att undersökningarna behandlar olika språk. För fem år sedan fanns troligtvis inga produkter alls som hanterade svenska – idag finns det åtminstone ett par stycken. Om man gjorde en motsvarande test med engelskspråkiga dokument idag skulle man troligtvis uppnå mycket bättre resultat än dem man fick 1995. Det har nämligen forskats åtskilligt inom detta område sedan 1995 och många nya produkter har lanserats.

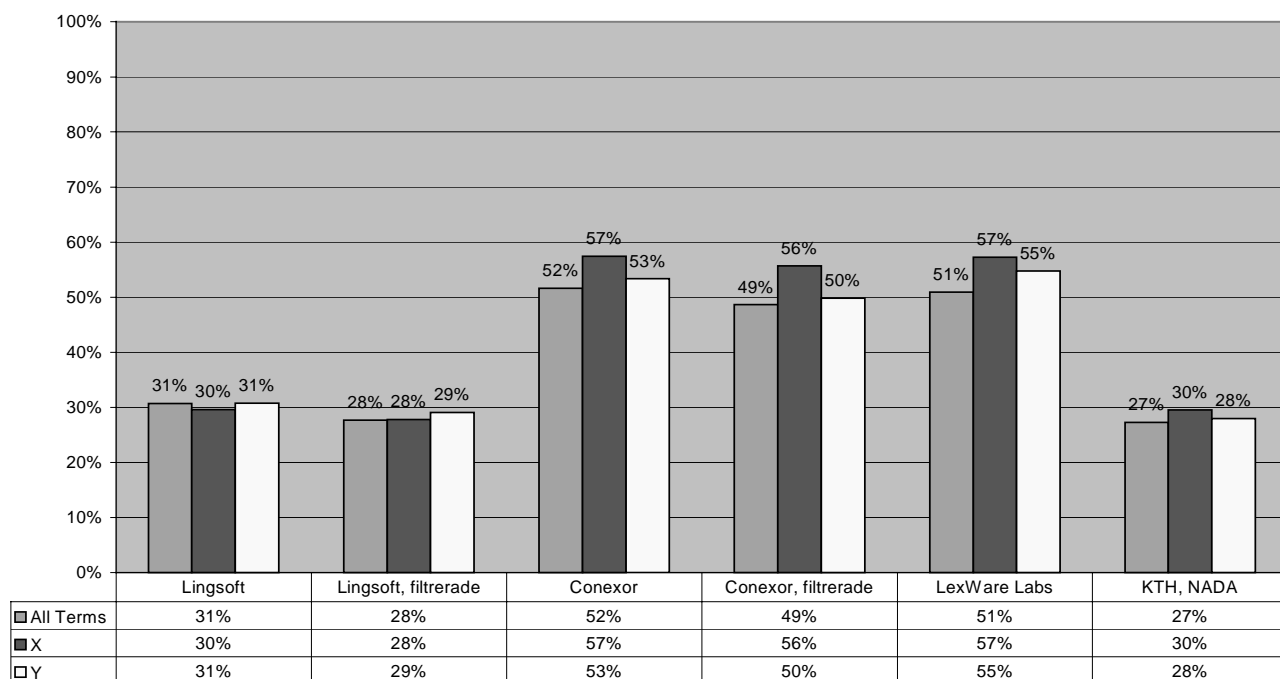


Diagram 11: Recall, jämförelse mellan kandidaterna

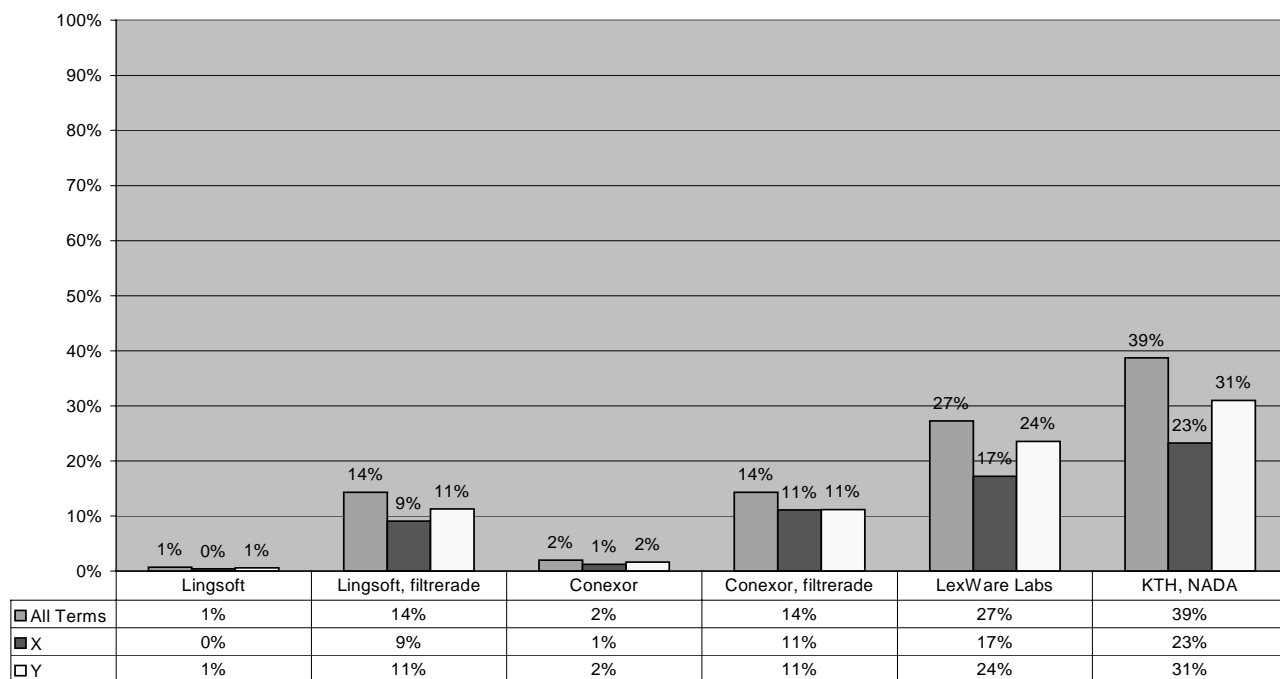


Diagram 12: Precision, jämförelse mellan kandidaterna

Av de siffror som framgår av diagram elva och tolv uppvisar LexWare Labs och KTH:s program bäst resultat. KTH uppvisar högre precision än LexWare Labs, men det beror till stor del på det faktum att deras resultatlista aldrig består av fler än fem termer. LexWare Labs överträffar vida KTH:s resultat när det gäller recall. Detta beror på att LexWare Labs program

är flexibelt så till vida att det plockar fram fler termer för ett längre dokument som indexerarna tilldelat många termer och färre för korta dokument, medan KTH:s program är statistiskt i det att det alltid plockar fram fem termer. Conexor uppnår lika bra resultat som Lex Ware Labs när det gäller recall, men priset för Conexors höga recallvärden är att värdena för precision är mycket låga.

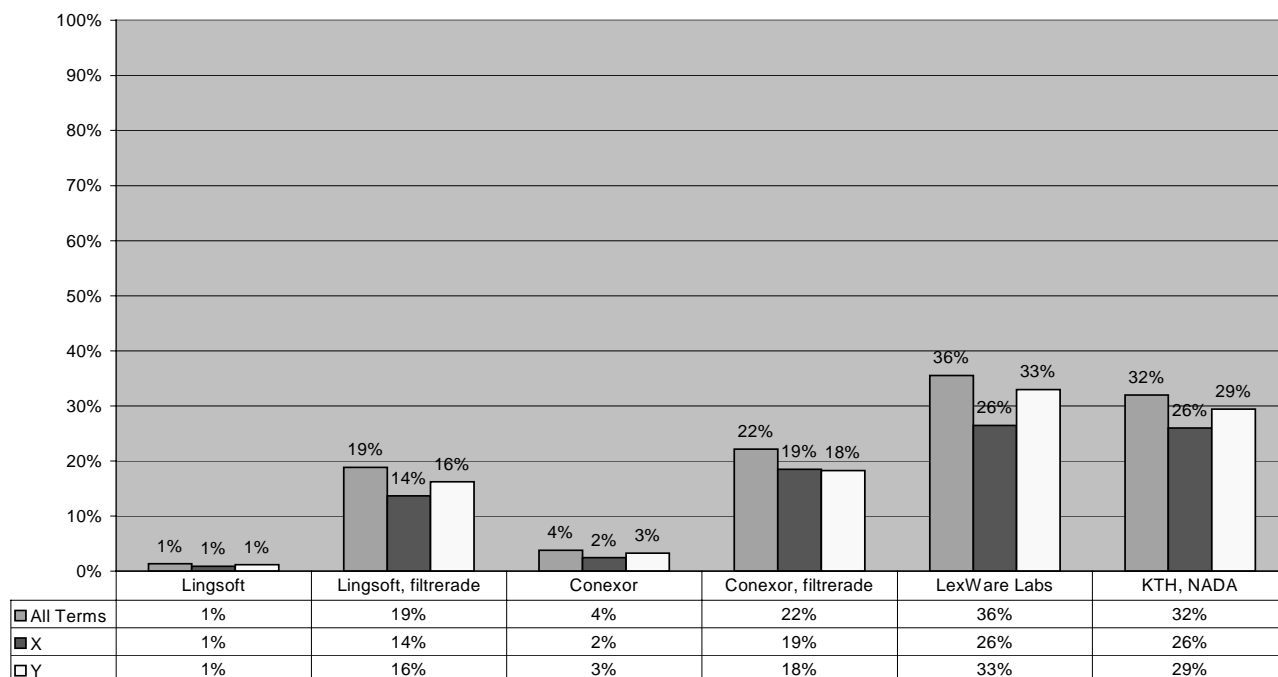


Diagram 13: F-värde, jämförelse mellan kandidaterna

Ser man till F-värdet, som ju är en sammanvägning av recall- och precisionvärdena så lyckas LexWare Labs program något bättre än KTH:s. LexWare Labs program får F-värden på mellan 26 % och 36 %, medan KTH:s program får F-värden på mellan 26 % och 32 %.

De värden som refereras till både i det här resultatkapitlet och i varje kandidats resultatkapitel är medelvärden. Det bör påpekas att resultaten kan variera mycket från ett dokument till ett annat. Vissa dokument har indexerats bra av programmen, medan andra dokument ger en nollprocentig indexeringskonsistens mellan indexerarna och programmen. För att kunna presentera resultaten på ett någorlunda överskådligt sätt var det dock nödvändigt att använda sig av medelvärden.

Förutom vad gäller Conexors recallresultat uppvisar Conexors och Lingsofts program resultat som ligger långt under de båda andra kandidaternas. Detta beror till stor del på att programmen plockar fram väldigt många termer för varje dokument. Men även de resultat som bygger på deras filtrerade term mängd är mycket lägre än LexWare Labs och KTH:s. Det bör framhållas att programmen inte är anpassade speciellt för den uppgift de testats på. Men man kan konstatera att Conexors och Lingsofts program så som de föreligger idag inte är mogna för att användas för automatisk indexering och troligen inte heller för datorstödd indexering.

Hur bra resultat ett program bör uppnå för att kunna användas för automatisk indexering är naturligtvis en bedömningsfråga från tillämpning till tillämpning och beror på vad man har för krav på indexeringen och hur det indexerade materialet är tänkt att användas. Vissa organisationer skulle kanske kunna nöja sig med den indexering som kan åstadkommas med hjälp av Lingsofts eller Conexors program, som i princip inte uppvisar någon överensstämmelse med indexerarna. I andra tillämpningar kanske man kräver en närapå hundra procentig överensstämmelse med manuell indexering. I riksdagens fall torde det krävas en relativt stor överensstämmelse, åtminstone för den indexering som är tänkt att bilda index i Volym G (registret) i riksdagstrycket. Det kan nämnas att i Europaparlamentets undersökning (Bureau van Dijk 1995:94) rekommenderas en indexeringskonsistens på åtminstone 50 % mellan ett programs output och en människas indexering för att programmet ska kunna komma i fråga. Då Europaparlamentet är en institution med likartade dokument och verksamhet som riksdagen kanske 50 % kan användas som en miniminivå här också. I sådant fall kan inget av de program som testats i den här undersökningen komma i fråga för automatisk indexering i det skick de befinner sig idag, eftersom inget av programmen presterar ett resultat som ens ligger i närheten av en indexeringskonsistens på 50 %.

Däremot är det fullt tänkbart att åtminstone LexWare Labs och kanske även KTH:s program skulle kunna användas för datorstött indexering. Den datorstödda indexeringen kunde rimligtvis gå till så att programmet processade ett dokument och därefter föreslog vissa termer för en indexerare, som efter att ha ögnat igenom dokumentet kunde godkänna eller förkasta dem, samt eventuellt lägga till ytterligare termer. Detta arbetssätt torde vara tidsbesparande jämfört med dagens manuella indexering. Samtidigt låter man inte ett program ta över helt, utan har fortfarande kvar indexerarens slutliga kontroll över indexeringen.

För en annan indexering än för index i Volym G i riksdagstrycket skulle alla programmens output kunna vara intressant. Resultatlistorna speglar ju väl dokumentens innehåll på ett innehållsmässigt mycket nära sätt.

7 Sammanfattning – diskussion

Målen för det arbete som redovisas i föreliggande rapport var dels att göra en marknadsundersökning över produkter för automatisk indexering, dels att testa vissa av de produkter som finns på marknaden. Testerna syftade till att se hur bra indexering man kan uppnå med hjälp av befintliga indexeringsprogram.

Marknadsundersökningen ledde till att en mängd företag, forskningsorganisationer och program kartlades och gav dessutom en förståelse för hur expansivt och dynamiskt området informationssökning, och därmed automatisk indexering, är.

Testerna ledde till närmare kontakt med fyra företag och erfarenhet av deras program. Resultaten av dessa tester har visat att inget av programmen, i det skick de idag föreligger, på ett helautomatiskt sätt kan ersätta den manuella indexering som görs vid Riksdagsbiblioteket. Där emot skulle åtminstone LexWare Labs program, och kanske även KTH:s, kunna vara alternativ när det gäller datorstödd indexering av riksdagstrycket.

Indexering är en bibliografisk verksamhet som syftar till att på ett objektivt sätt skapa länkar mellan en informationssökare och en dokumentmängd. När det är människor som utför indexeringen är det dock oundvikligt att resultatet blir en smula subjektivt. Indexeringen uppvisar större variation ju fler indexerare som är inblandade, men även en enskild indexerare varierar i sin indexering över tiden. Att använda sig av indexerares val av termer som facit för en utvärdering av indexeringsprogram kan således ifrågasättas. Med detta resonemang i åtanke och för att i någon mån undvika för mycket godtycklighet i undersökningen har programmens output jämförts med både indexerarnas sammanlagda termmängd och indexerare X:s och indexerare Y:s termmängder var och en för sig.

Indexering är inget självändamål utan syftar till att informationssökare ska kunna återfinna dokument med hjälp av de termer som tilldelats dokumenten. Det som egentligen borde undersökas är således hur bra man kan återfinna dokument med hjälp av de termer som ett program tilldelat dokumenten versus de termer som indexeraren tilldelat. Den undersökning som redovisas i denna rapport bygger på antagandet att dokument i hög grad återfinns med hjälp av de termer som indexerarna tilldelar dokumenten. Indexerarnas val av termer har således utgjort facit för resultatlistorna av de program som testats. Om mer tid och resurser hade funnits skulle det ha varit mycket intressant att jämföra hur bra man kan återfinna dokument med de olika metoderna, eftersom det är denna verksamhet som indexering av dokument syftar till.

Testmaterialet som användes i den ovan redovisade undersökningen bestod, som nämnts under 6.2, av 4 % av den totala dokumentmängden. Ett testmaterial som utgör en så liten andel av den totala dokumentmängden är egentligen för litet. Inom datorlingvistik är den rådande praxisen för jämförande undersökningar av den här typen att testmaterialet ska utgöra en tiondel av den totala dokumentmängden (Norgard och Plaunt 1997:9). Men eftersom vi ville ha möjlighet att kunna gå igenom resultaten för varje dokument manuellt var det viktigt att testkorpusen inte svällde till ohanterlighet. I och med önskemålet om att testmaterialet skulle avspegla den verkliga dokumentfördelningen mellan dokumenttyperna blev vissa av dem väldigt små i testmaterialet. De slutsatser som t.ex. dras av de resultat som sex propositioner ger kan inte sägas vara tillförlitliga utan måste ses enbart som en indikation på en viss tendens.

En intressant fortsättning och utvidgning av föreliggande undersökning vore att göra ett detaljstudium av de termer som föreslås av de testade indexeringsprogrammen. Vid en dylik undersökning borde man beakta de semantiska relationer som förekommer mellan de termer som föreslås av programmen och dem som föreslås av indexerarna. Om man kunde explicitgöra dessa relationer skulle kanske resultaten av jämförelserna mellan programmets output och indexerarnas indexering bli annorlunda än i den undersökning som redovisas i den här rapporten. En sådan utvidgning ligger dock utanför ramen för den här delstudien inom projektet Automatisk indexering.

De data som testerna gett är intressanta också ur andra aspekter än just jämförelsen mellan människa och maskin. Indexerarnas dubbelindexering skulle t.ex. kunna användas för att utvärdera och förbättra de indexeringsregler som tillämpas vid sektionen för indexering och registerproduktion. Programmets output ger lärdom om vilka termer som verkligen används av riksdagsledamöter (och regeringskansliet när det gäller propositionerna) och är en rik källa till information om språkbruket på riksdagen. Dessutom skulle programmets utdata kunna användas för att hitta nya termer vid en uppdatering av Riksdagens tesaurus.

Slutligen kan sägas att för att åstadkomma automatisk indexering krävs större och mer raffinerade system än dem som testats i den här delstudien. Sådana finns i dagens läge inte för svenska. Det skulle innebära ett väsentligt åtagande att göra en försvenskning av de system som skulle kunna klara av indexeringsuppgiften på ett någorlunda helautomatiskt sätt. På samma sätt innebär det mycket arbete att utveckla de program som testats i den här delstudien till storskaliga och robusta indexeringsystem.

Litteraturförteckning

Benito, Miguel, 1993, Bibliografisk kontroll, Taranco, Borås.

Bureau van Dijk, 1995, Parlement Européen, Evaluation des opérations pilotes d'indexation automatique (Convention spécifique n° 52556), Rapport d'évaluation finale.

Crystal, David, 1994, Rediscover Grammar, Longman Group UK Limited.

Dahl, Östen, 1982, Grammatik, Studentlitteratur.

Dura, Elzbieta, 1998, Parsing Words, Göteborgs universitet.

Hellsten, Unn & Rosfelt, Margareta, 1999, Ämnesordsindexering en handledning, Kungliga biblioteket, Stockholm.

Hjørland, Birger, 1993, Emnerepraesentation og informationssøgning: bidrag til en teori på kundskabsteoretisk grundlag, Borås:Valfrid.

Karlgren, Jussi, 2000, Stylistic Experiments for Information Retrieval, Stockholms universitet och Swedish Institute of Computer Science.

Koskenniemi, Kimmo, 1983, Two-Level Morphology. A General Computational Model of Word Form Recognition and Production, University of Helsinki: Publications of the Department of General Linguistics.

Källgren, Gunnel, 1984, Automatisk excerpering av substantiv ur löpande text. Ett möjligt hjälpmedel vid datoriserad indexering?, Stockholms universitet.

Lindkvist Michailaki, Elisabet, 1999, Indexeringsregler, Riksdagsbiblioteket, Sveriges Riksdag.

Ljung, Magnus och Ohlander, Sölve, 1982, Allmän Grammatik, Gleerups.

LT TCR: Text Categorization and Text Routing, 990906,
<http://www.ltg.ed.ac.uk/software/tcr>.

Luhn, H.P., 1957, A Statistical Approach to Mechanical Encoding and Searching of Literary Information, IBM Journal of Research and Development.

Luhn, H.P., 1958, The Automatic Creation of Literature Abstracts, IBM Journal of Research and Development.

Luhn, H.P., 1959, Auto-Encoding of Documents for Information Retrieval Systems, Modern Trends in Documentation, Pergamon Press, London.

Machine Aided Indexer (M.A.I.), 990817, <http://www.dataharmony.com/do.html>.

Milstead, Jessica L., 1998, Use of Thesauri in the Full-Text Environment, 990727, <http://www.jelem.com/full.html>.

NEO, Nationalencyklopedins ordbok, 1996, Bokförlaget Bra Böcker, Höganäs.

Norgard, Barbara A. och Christian Plaunt, 1997, An Association Based Method for Automatic Indexing with a Controlled Vocabulary.

Practical Handbook of Genetic Algorithms, 000331, <http://www.crcpress.com/index.htm?catalog/2529>.

Riksdagens tesaurus, 1997, Riksdagens dokumentenhet, Sveriges Riksdag.

Salton, G., Wong, A., och Yang, C.S., 1975, A Vector Space Model for Automatic Indexing, Communications of the ACM.

Sparck Jones, Karen och Willet, Peter, 1997, Readings in Information Retrieval, Morgan Kaufmann Publishers, Inc., San Fransisco, California.

Sproat, R, 1992, Morphology and Computation, MIT Press.

Thesaurus Construction, 990707, <http://instruct.uwo.ca/gplis/677/thesaur/main00.html>.

Turney, Peter, 1997, Extraction of Keyphrases from Text: Evaluation of Four Algorithms, National Research Council Canada, Institute for Information Technology.

Turney, Peter, 1999, Learning to Extract Keyphrases from Text, National Research Council Canada, Institute for Information Technology.

Turtle, H. och Croft, W.B., 1990, Inference networks for document retrieval, Proceedings of the 13th International Conference on Research and Development in Information Retrieval, New York:Association for Computing Machinery.

Subject Indexing: Principles and Practices in the 90's. Proceedings of the IFLA satellite Meeting held in Lisbon, Portugal, 17-18 August 1993., UBCIM Publications – New Series, Vol. 15, 1995, Edited by Holley, Robert P., Mc Garry, Dorothy, Duncan, Donna, Svenonius, Elaine, Saur, K.G., München New Providence. London. Paris.

van Rijsbergen, C.J., 1979, Information Retrieval. Second edition. London: Butterworths.

What is an Artificial Neural Network?, 000331, <http://www.emsl.pnl.gov:2080/proj/neuron/neural/what.html>.

Bilaga Lingsoft

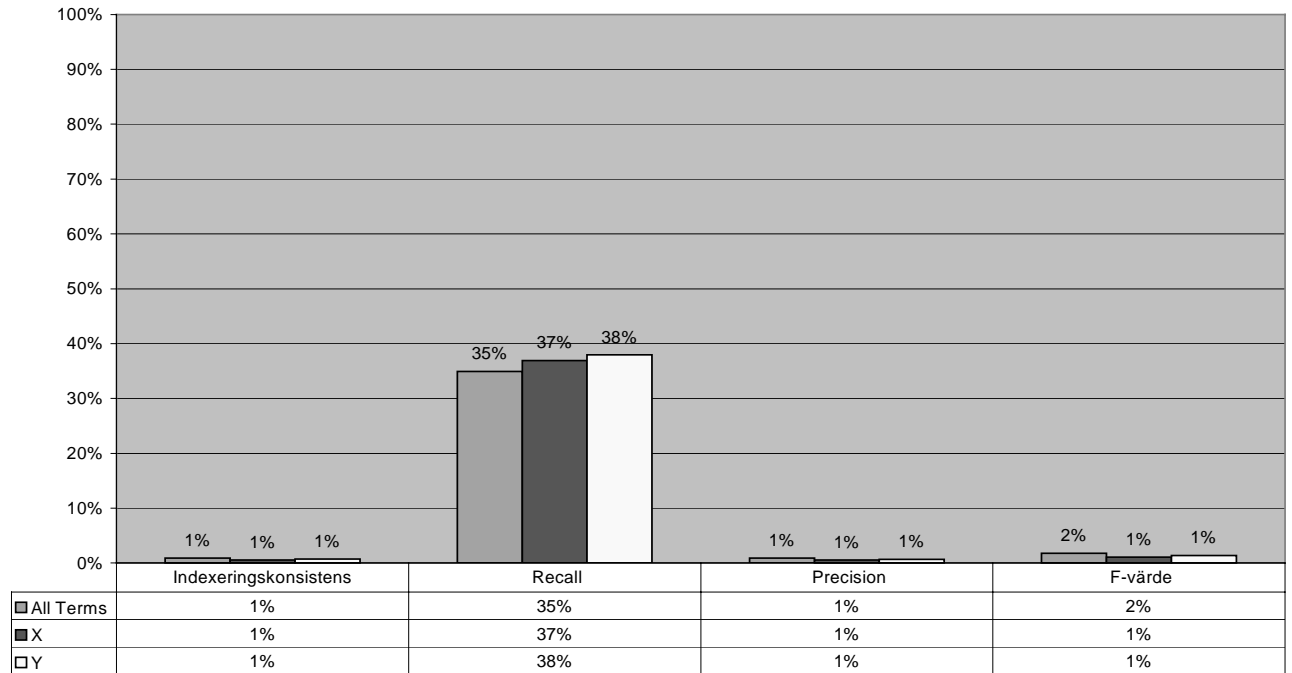


Diagram 14: Allmänna motioner, Lingsofts alla termer och indexerarnas termer

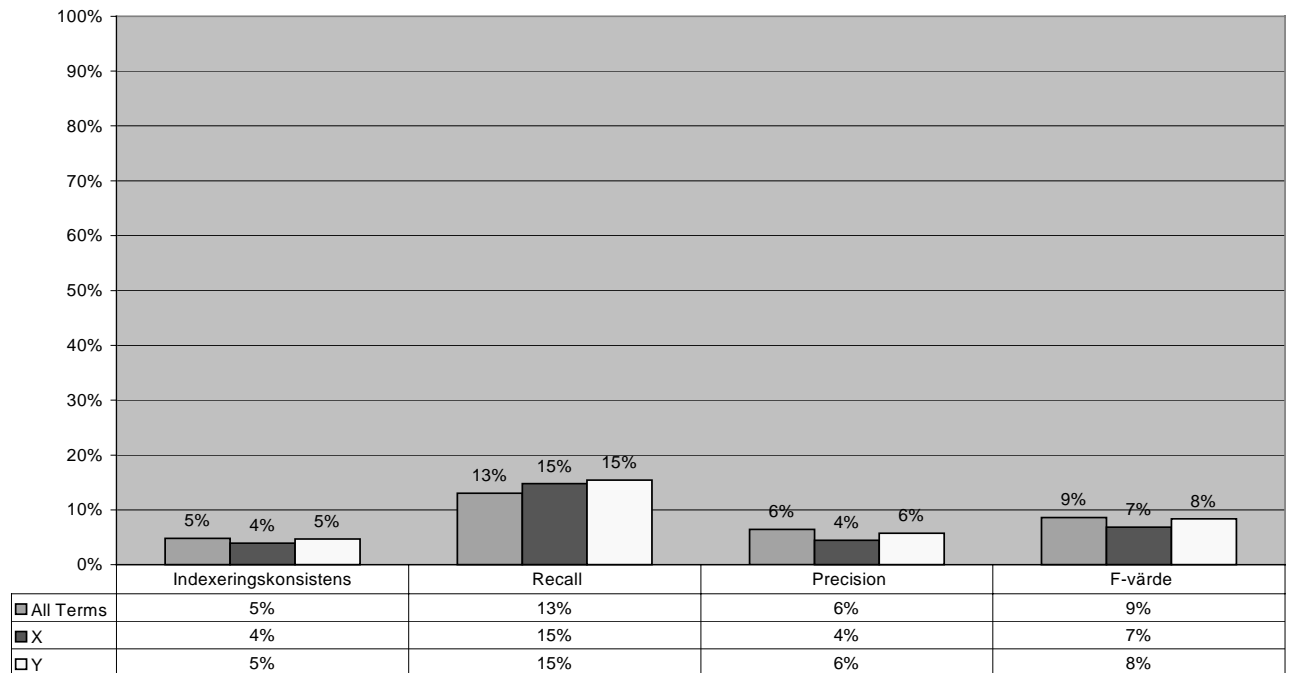


Diagram 15: Allmänna motioner, Lingsofts tio högst rankade termer och indexerarnas termer

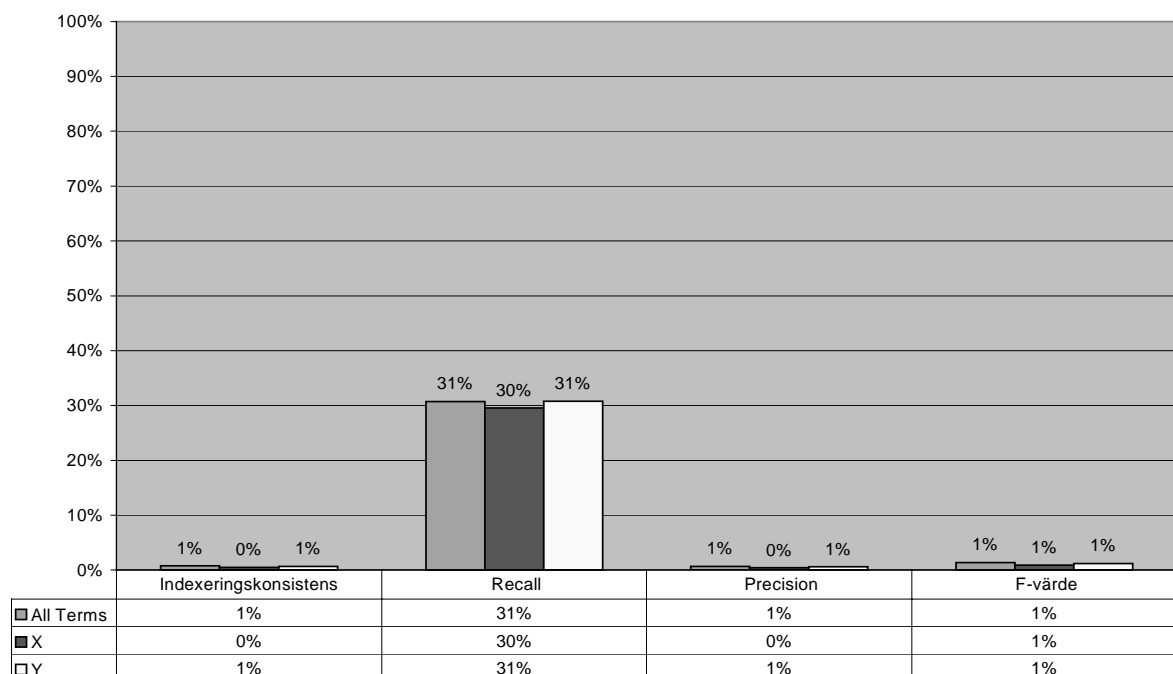


Diagram 16: Följdmotioner, Lingsofts alla termer och indexerarnas termer

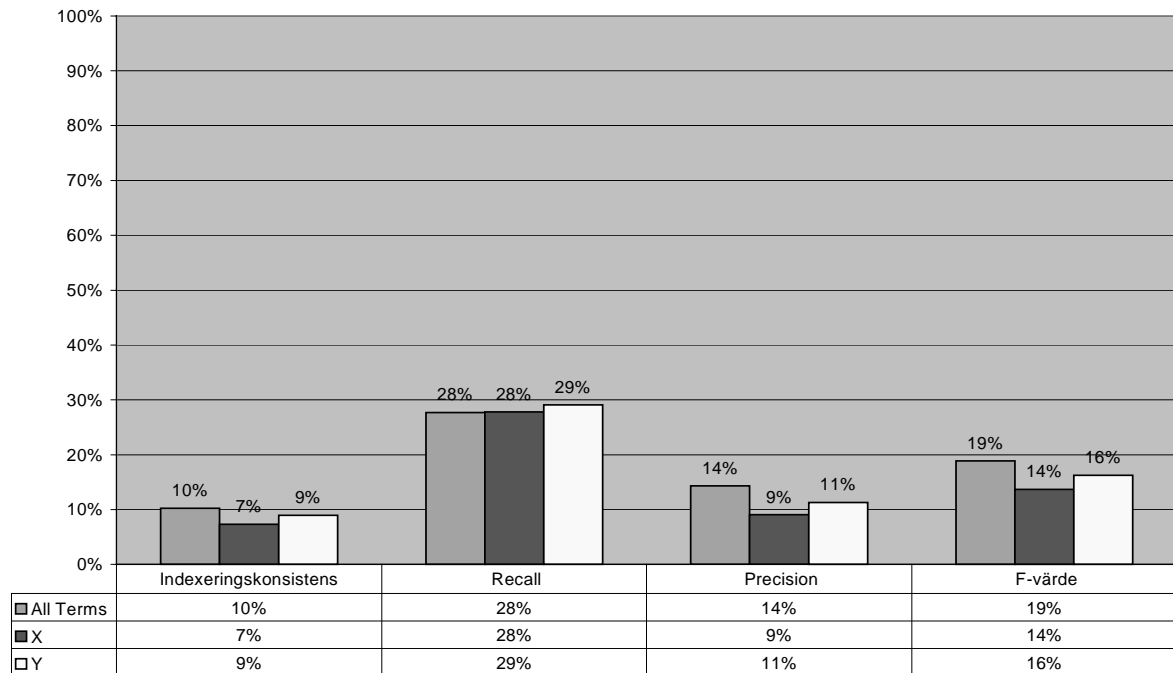


Diagram 17: Följdmotioner, Lingsofts filtrerade termer och indexerarnas termer

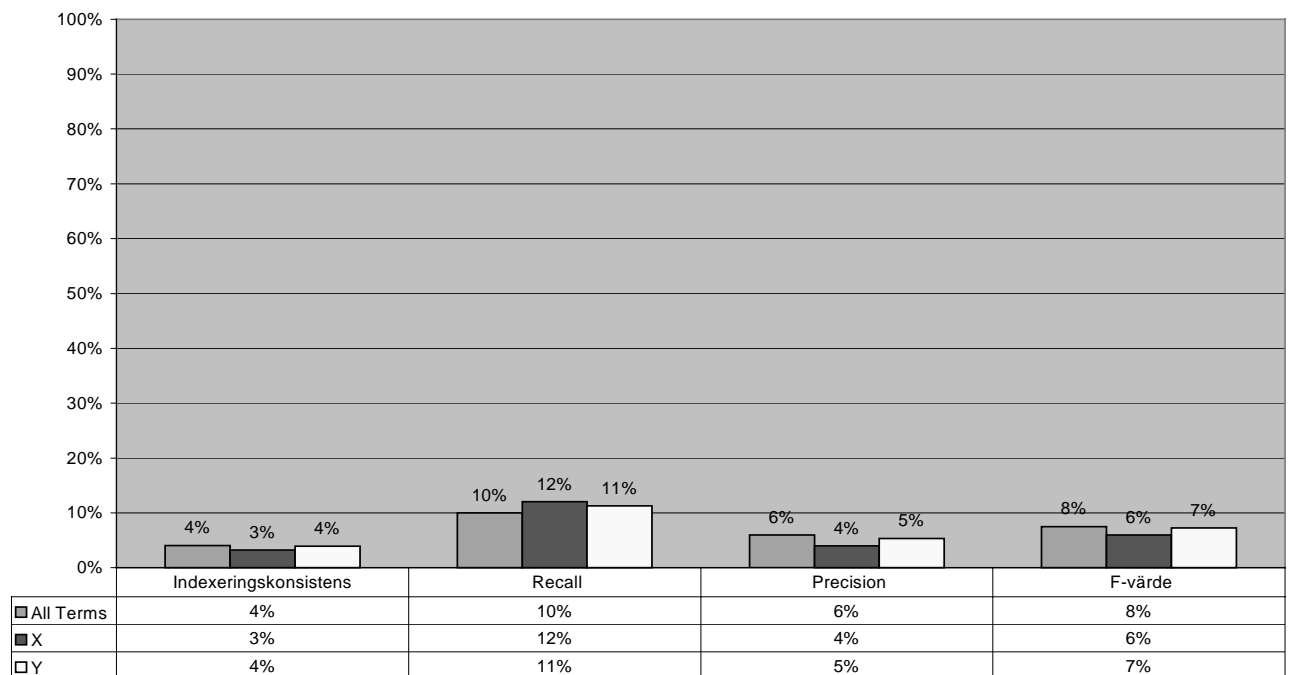


Diagram 18: Följdmotioner, Lingsofts tio högst rankade termer och indexerarnas termer

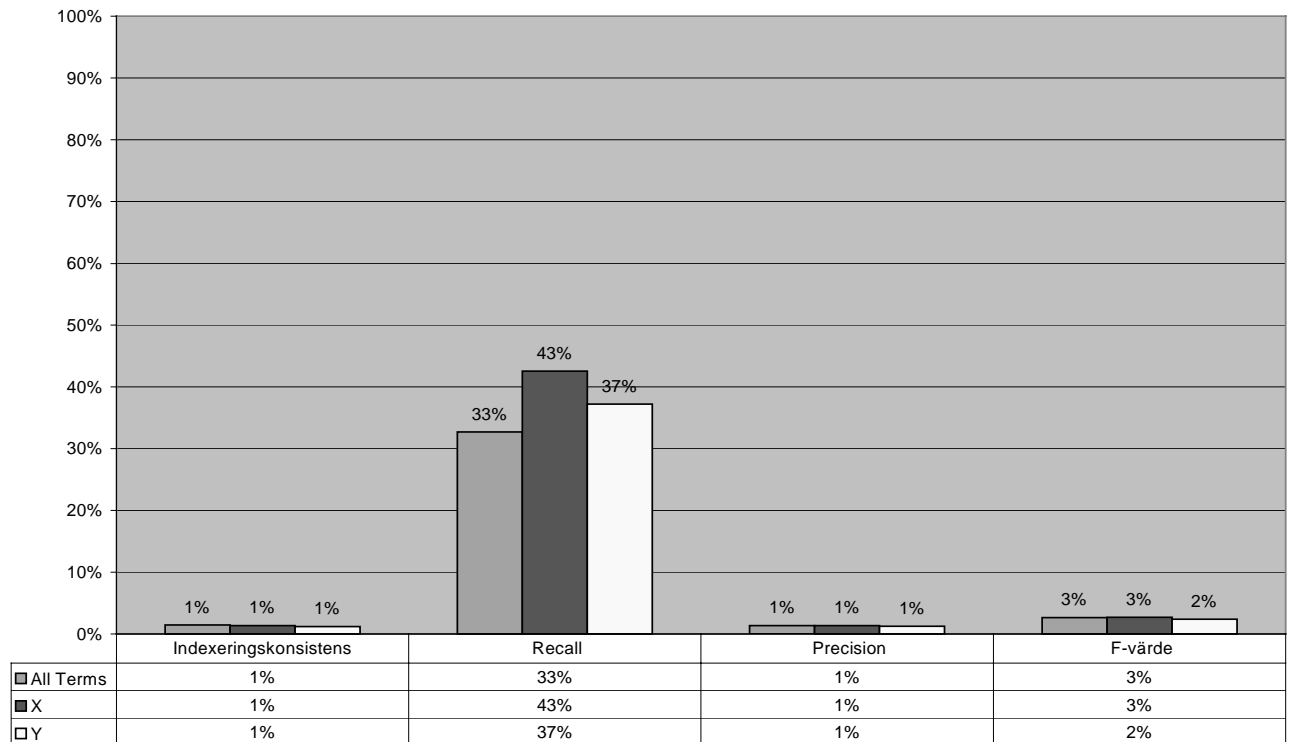


Diagram 19: Interpellationer, Lingsofts alla termer och indexerarnas termer

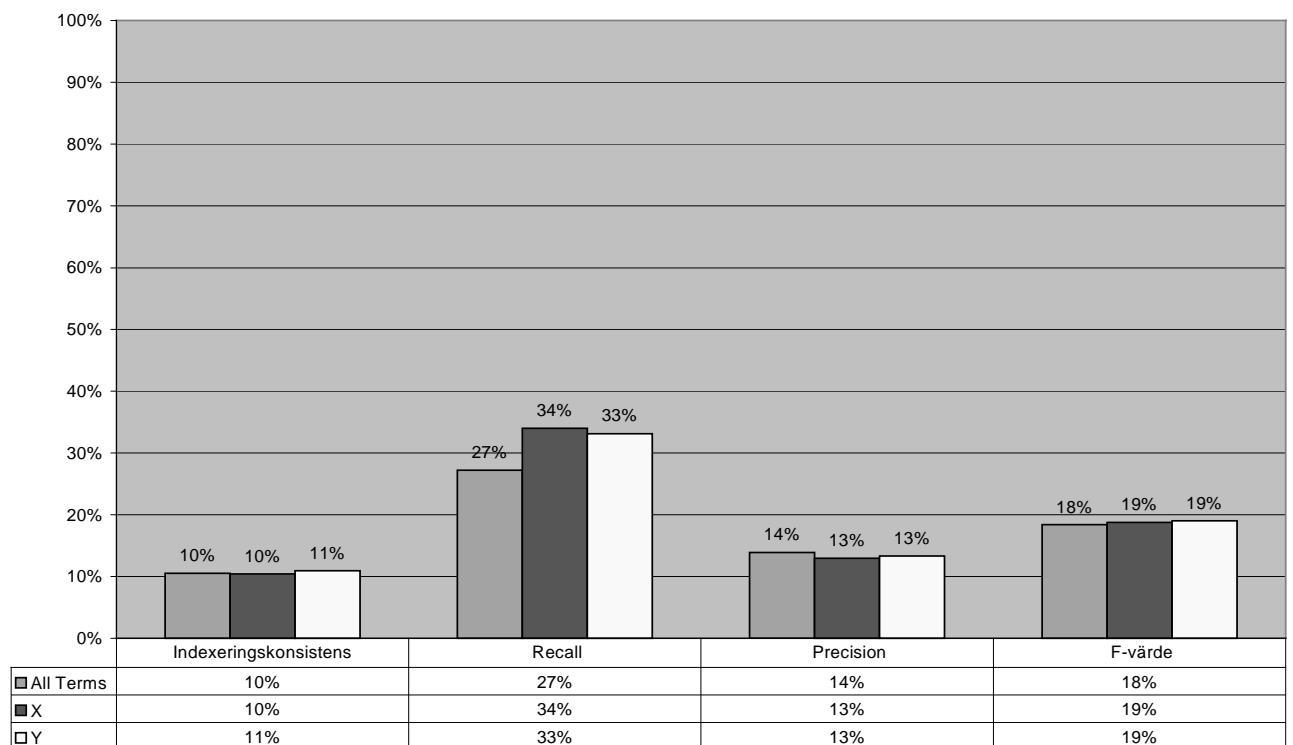


Diagram 20: Interpellationer, Lingsofts filtrerade termer och indexerarnas termer

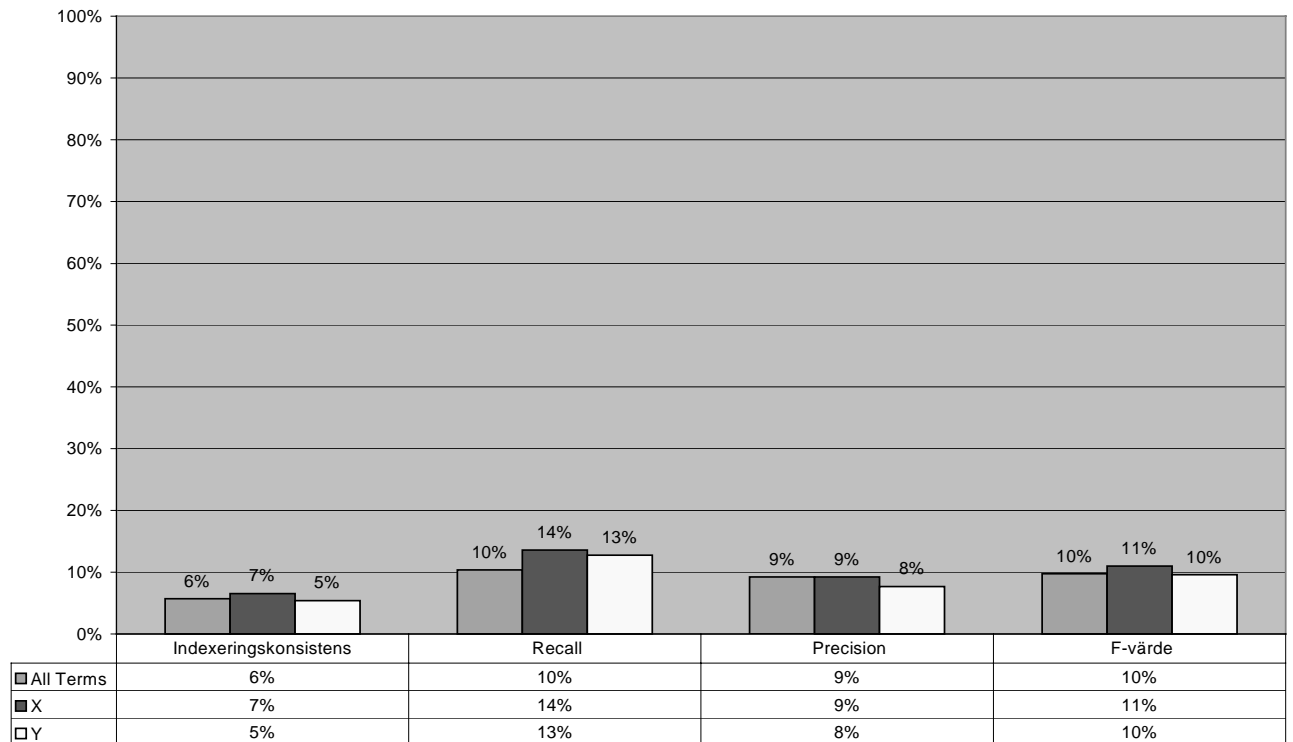


Diagram 21: Interpellationer, Lingsofts fem högst rankade termer och indexerarnas termer

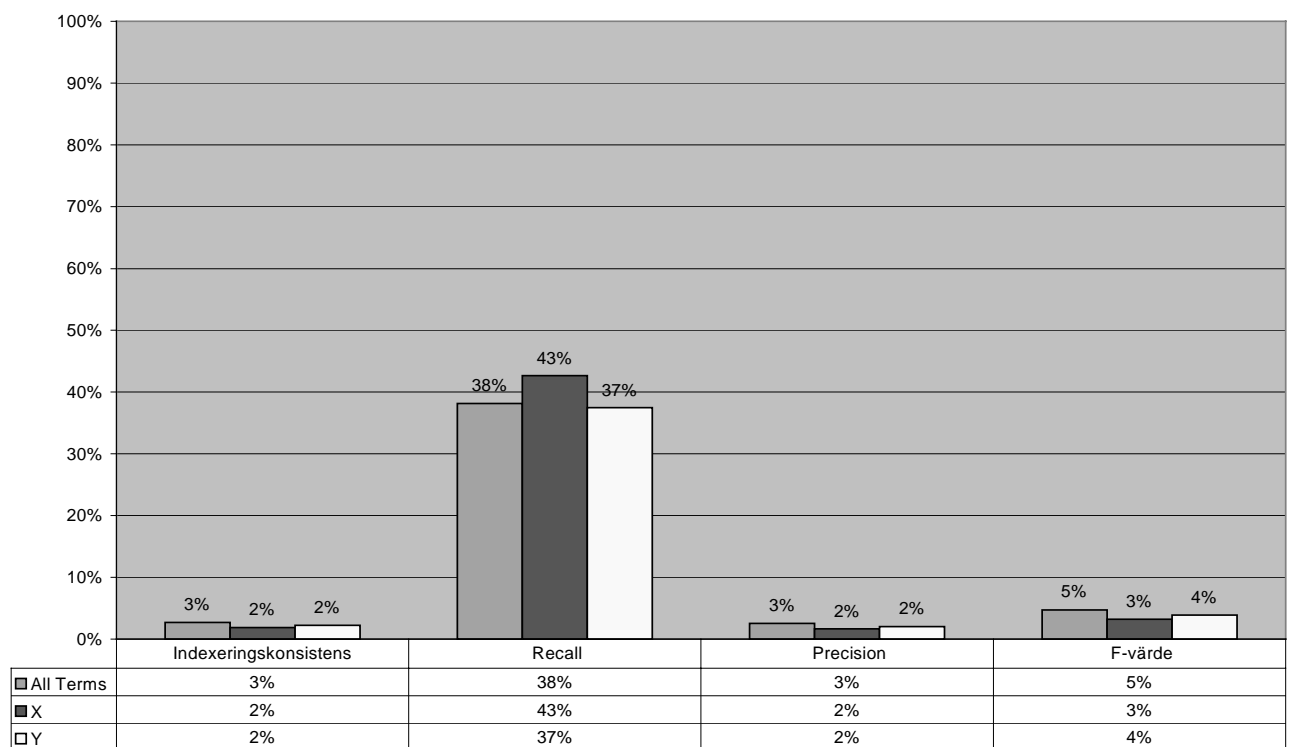


Diagram 22: Frågor, Lingsofts alla termer och indexerarnas termer

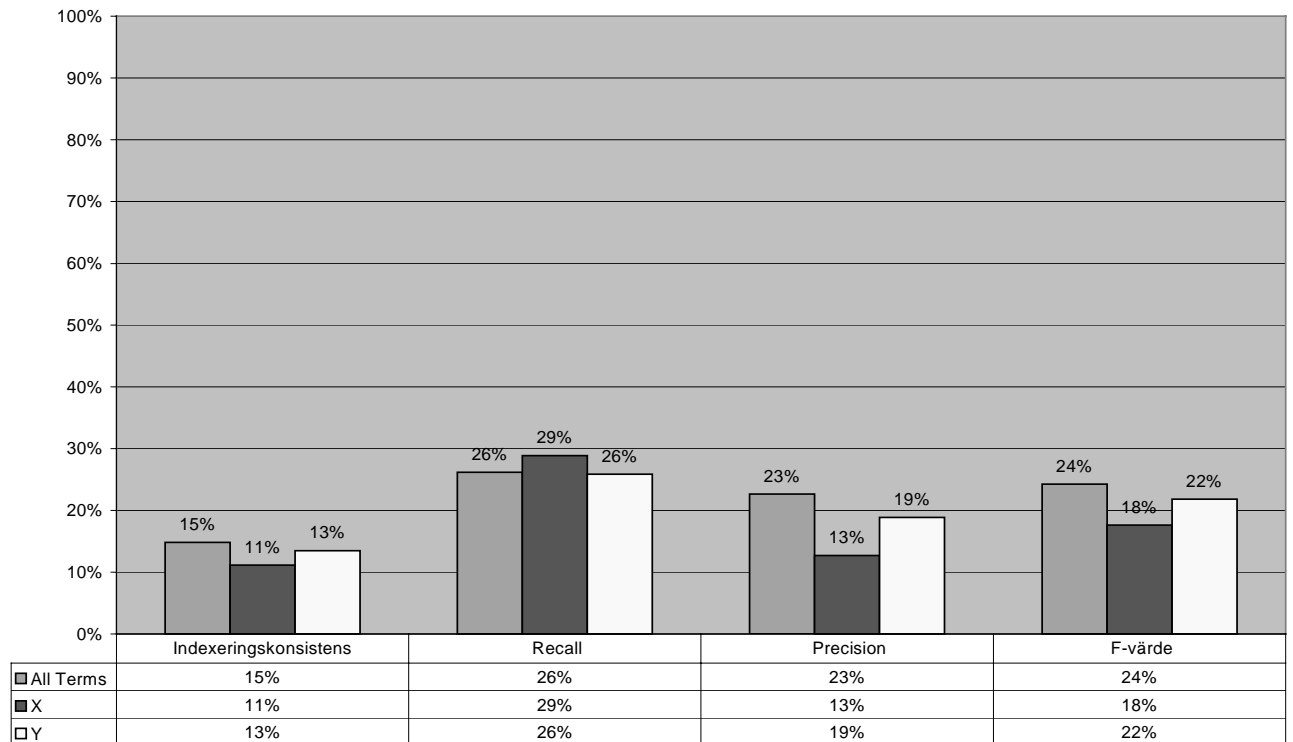


Diagram 23: Frågor, Lingsofts filtrerade termer och indexerarnas termer

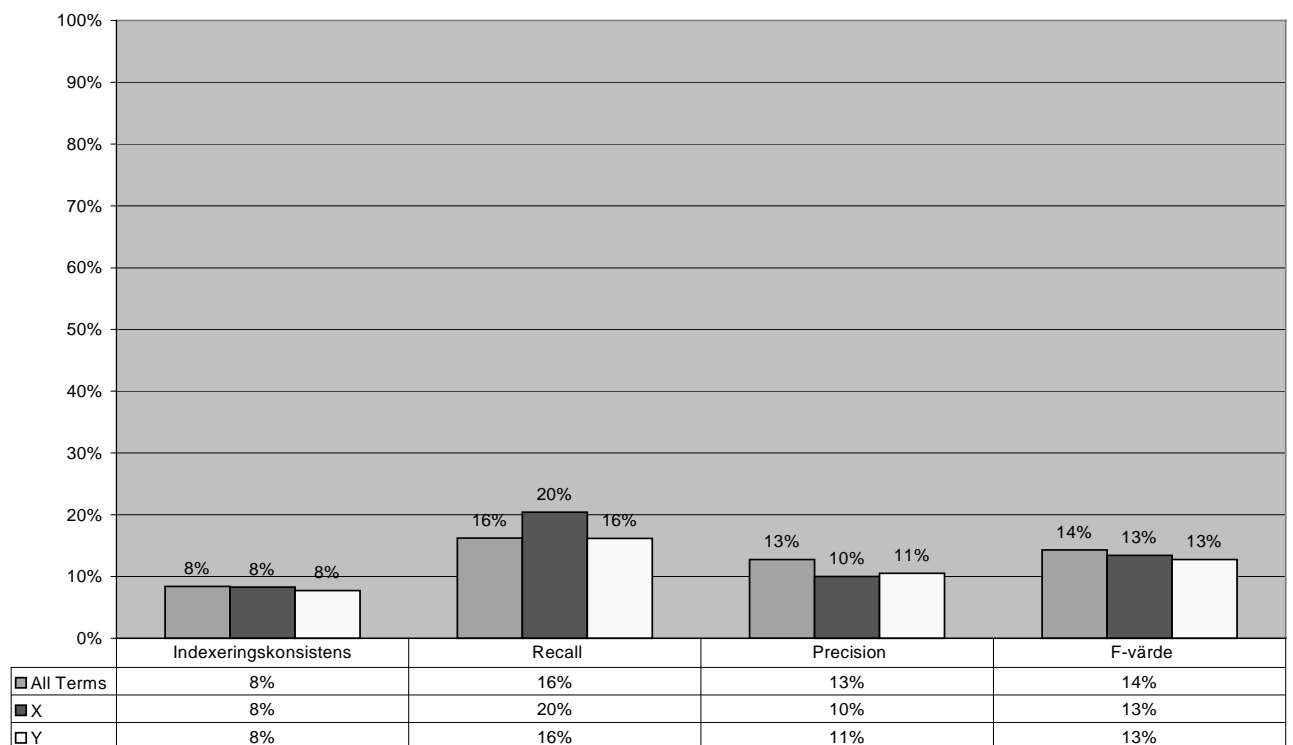


Diagram 24: Frågor, Lingsofts fem högst rankade termer och indexerarnas termer

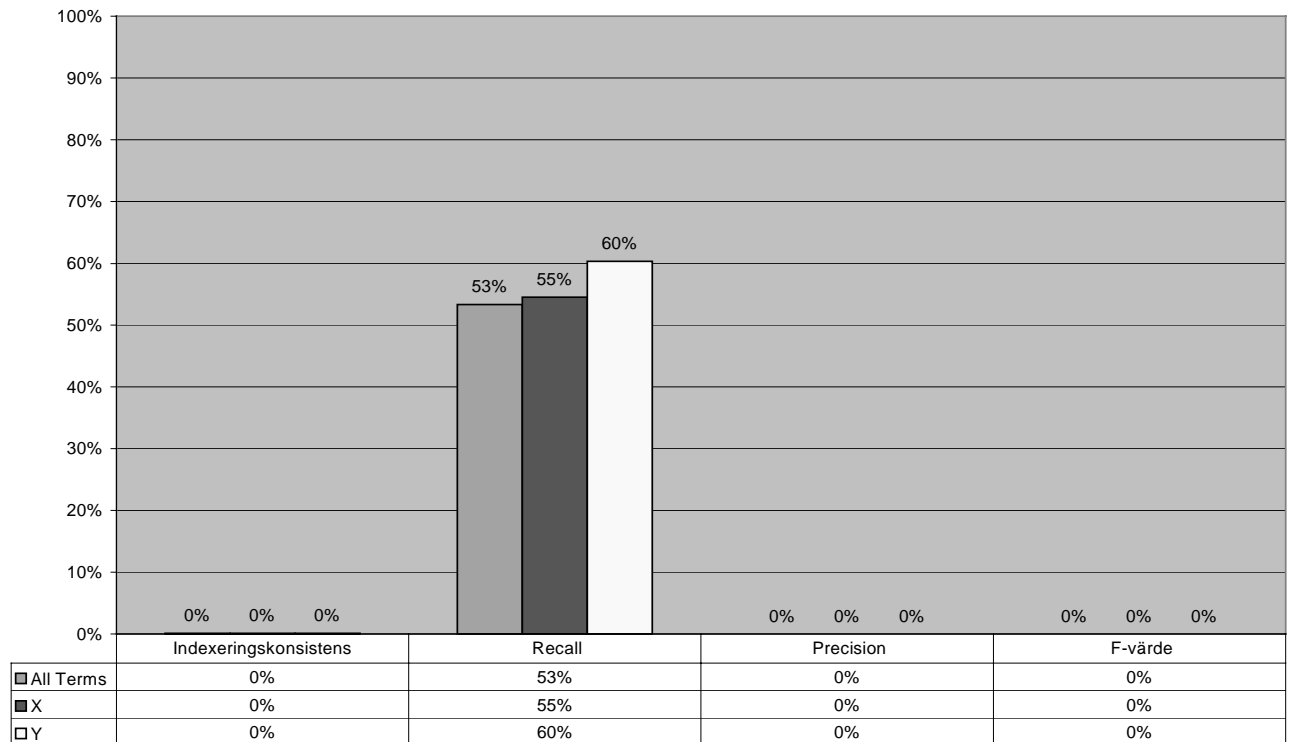


Diagram 25: Propositioner, Lingsofts alla termer och indexerarnas termer

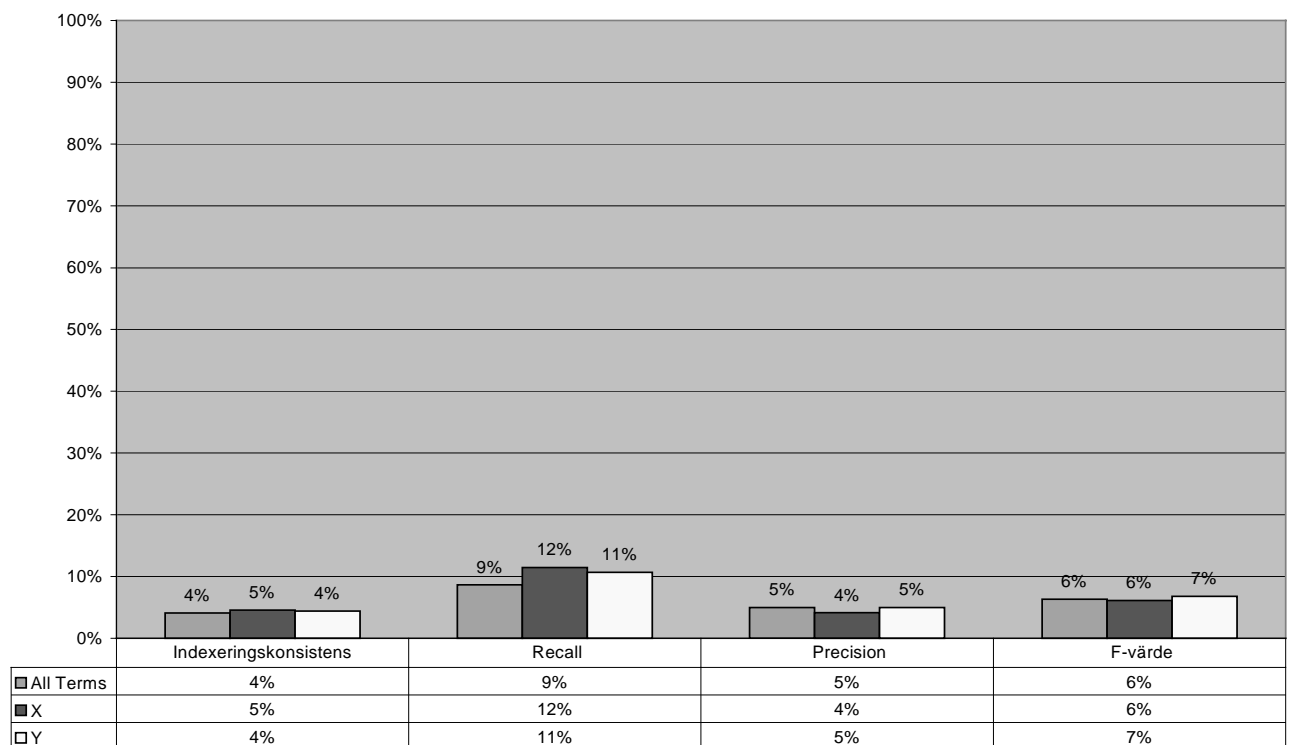


Diagram 26: Propositioner, Lingsofts tjugo högst rankade termer och indexerarnas termer

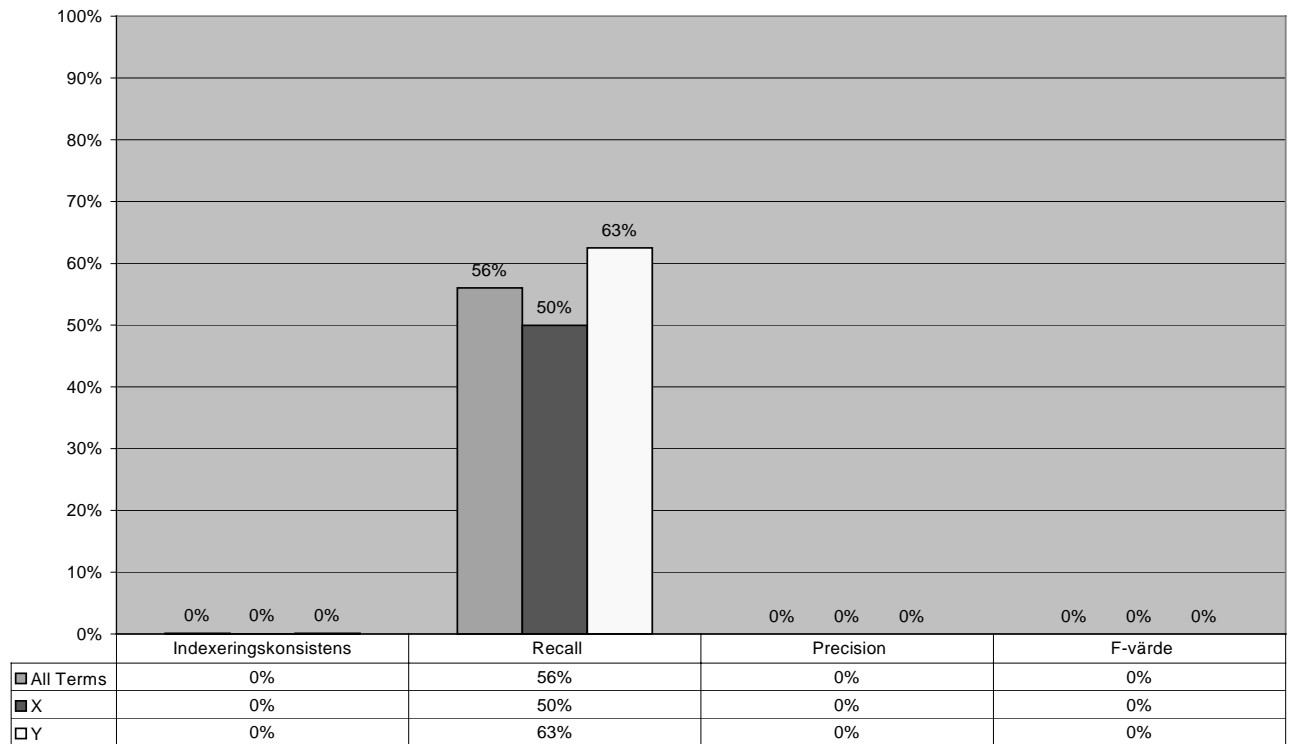


Diagram 27: Skrivelse, Lingsofts alla termer och indexerarnas termer

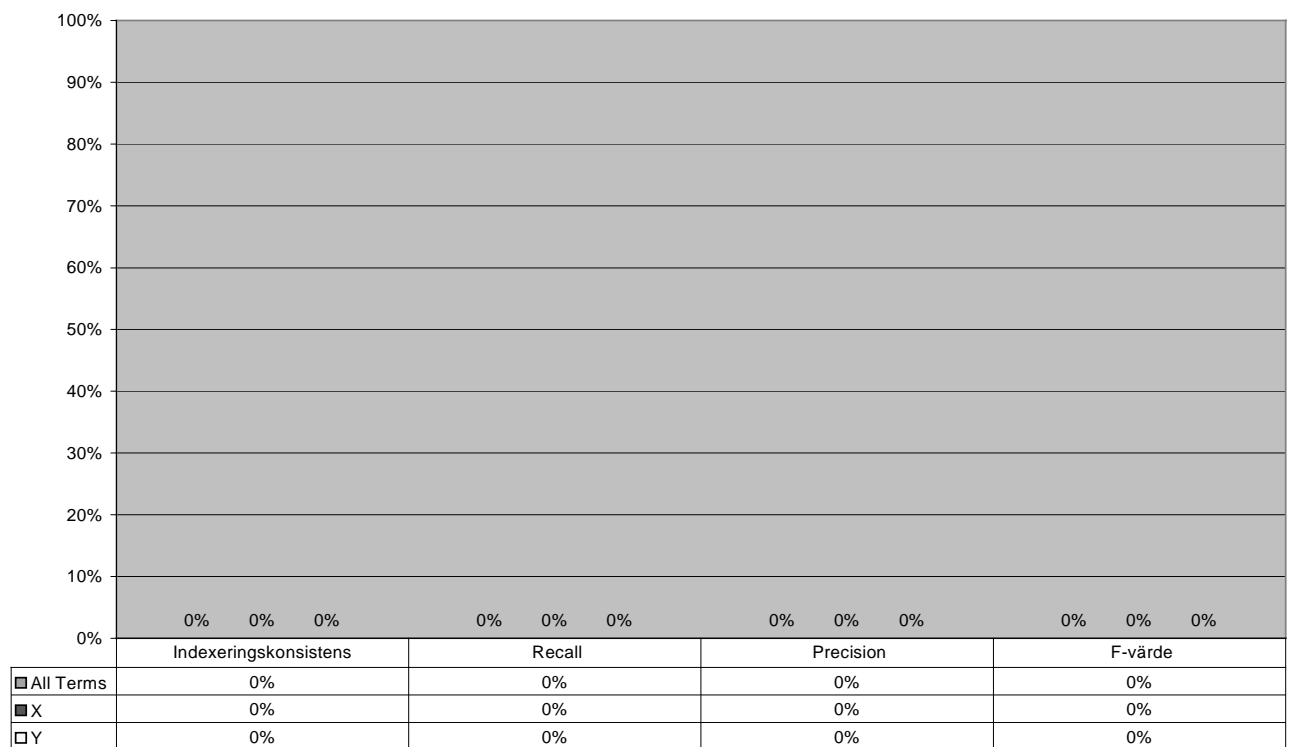


Diagram 28: Skrivelse, Lingsofts tjugo högst rankade termer och indexerarnas termer

Bilaga Conexor

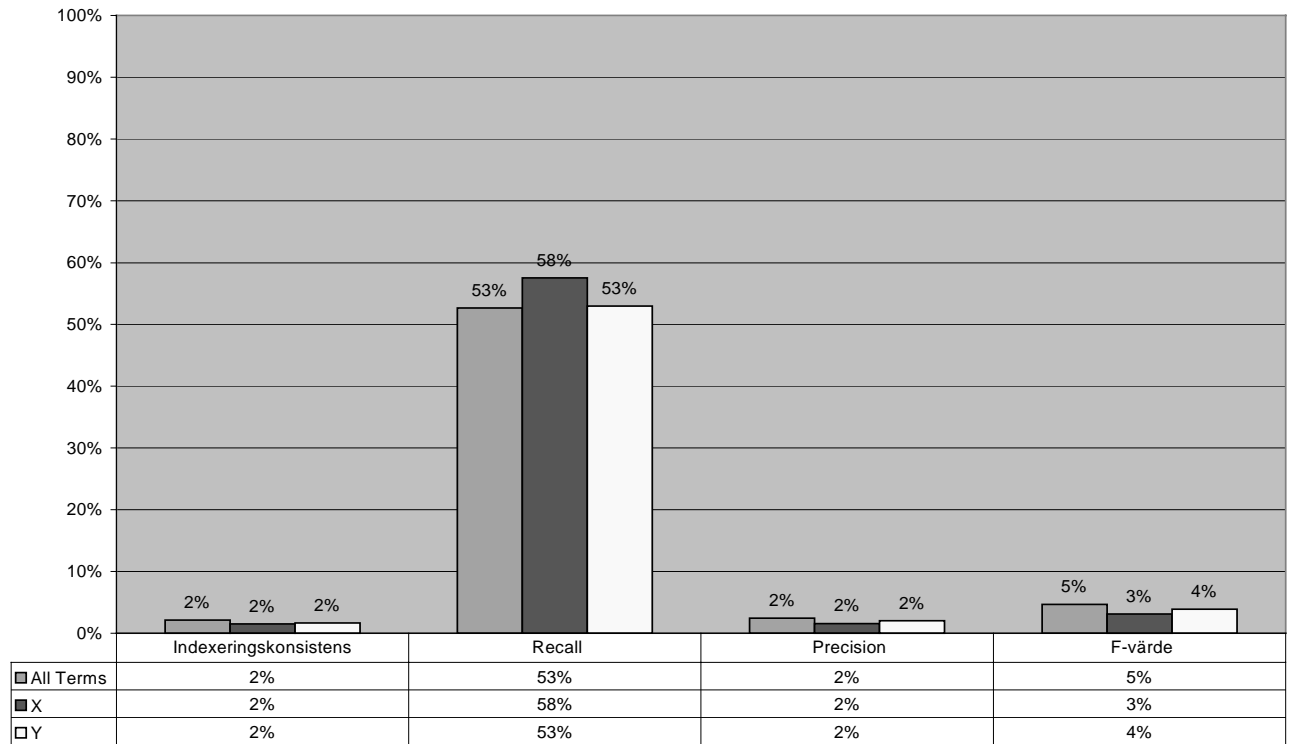


Diagram 29: Allmänna motioner, Conexors alla termer och indexerarnas termer

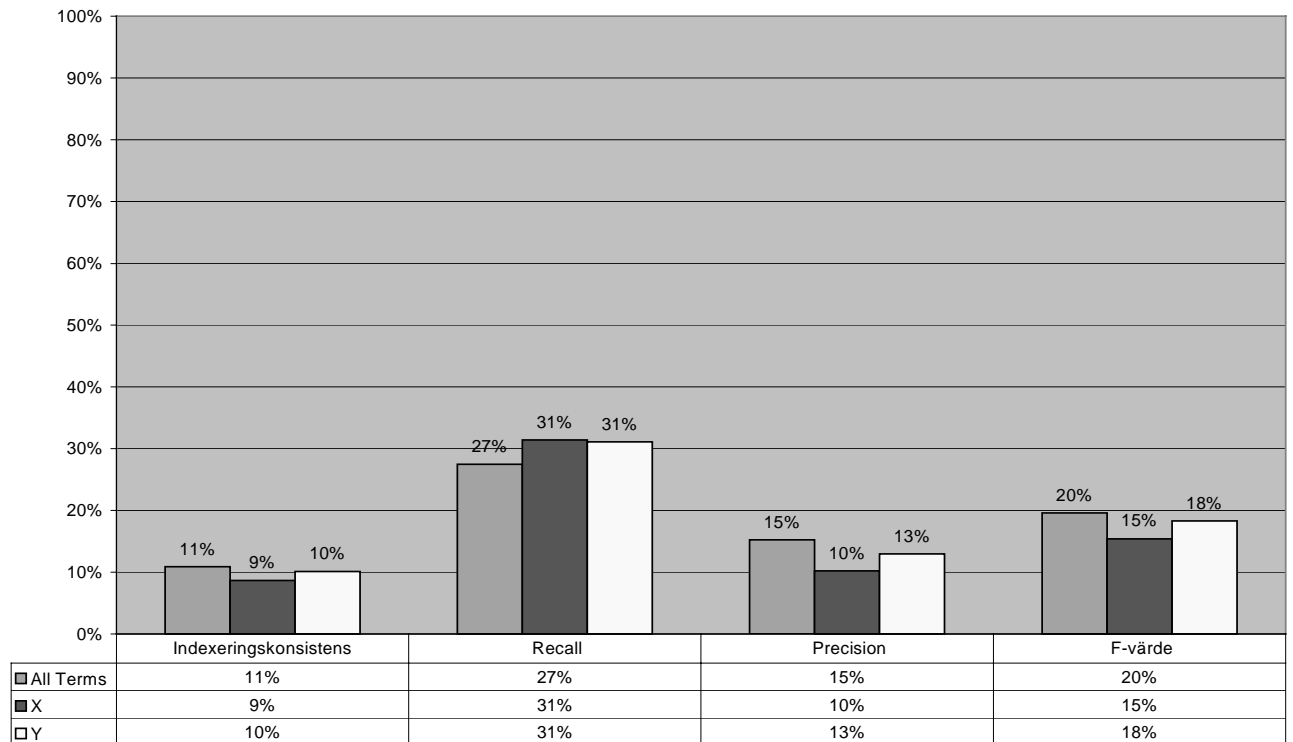


Diagram 30: Allmänna motioner, Conexors tio högst rankade termer och indexerarnas termer

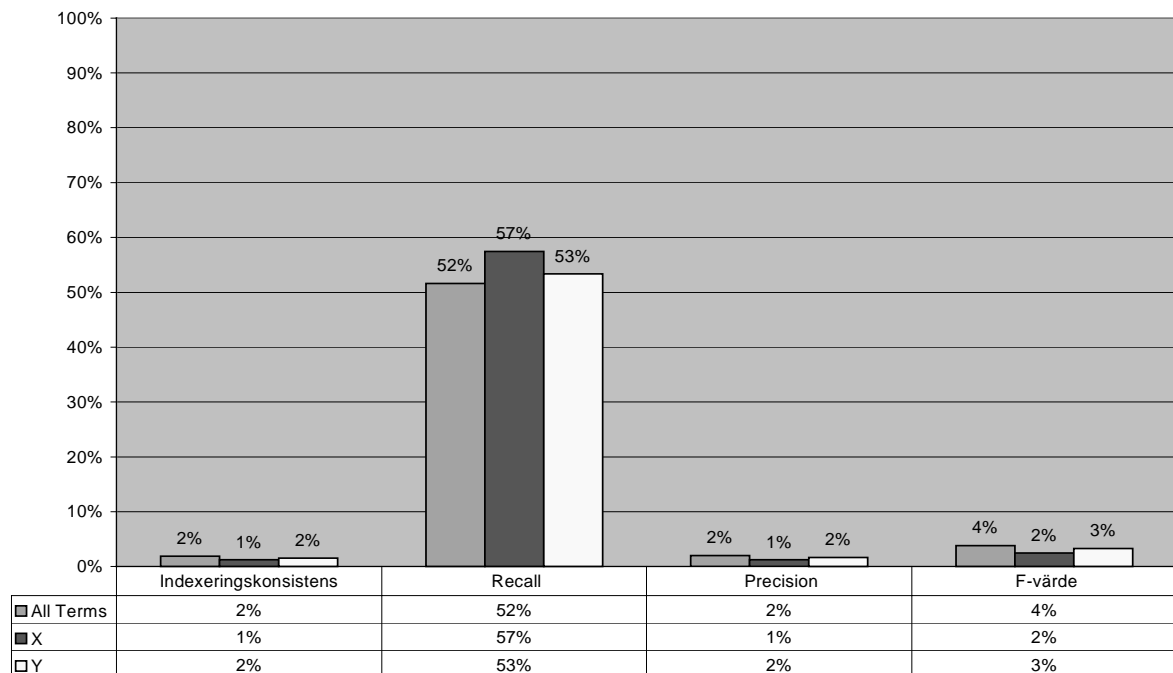


Diagram 31: Följdmotioner, Conexors alla termer och indexerarnas termer

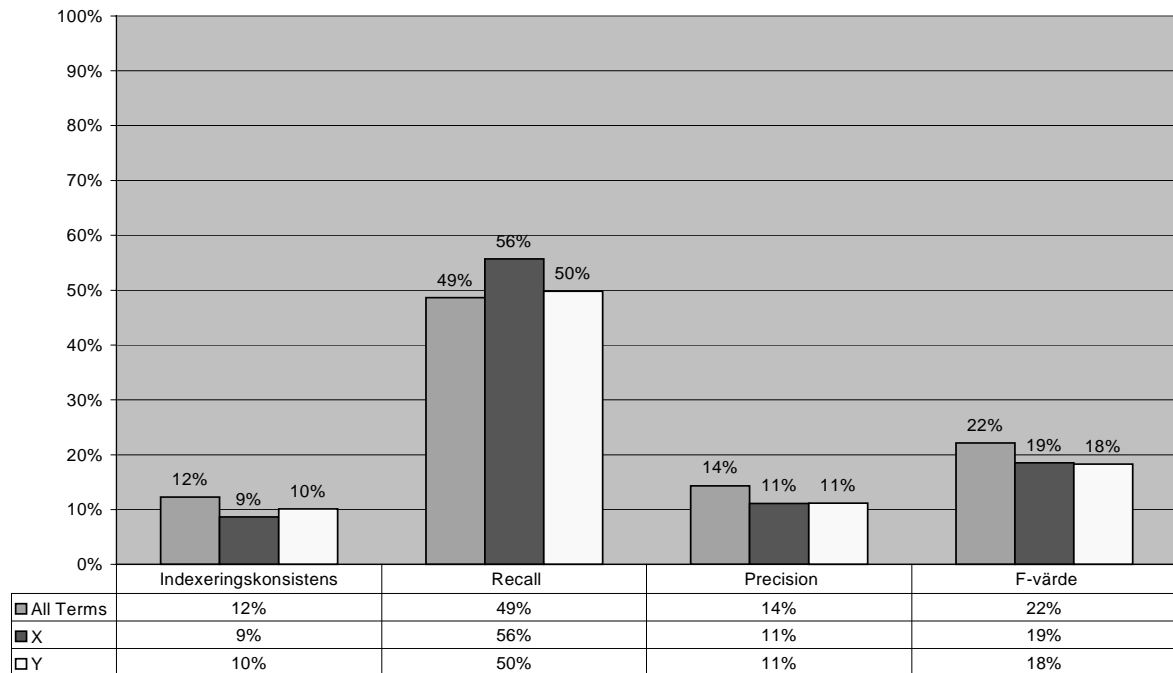


Diagram 32: Följdmotioner, Conexors filtrerade termer och indexerarnas termer

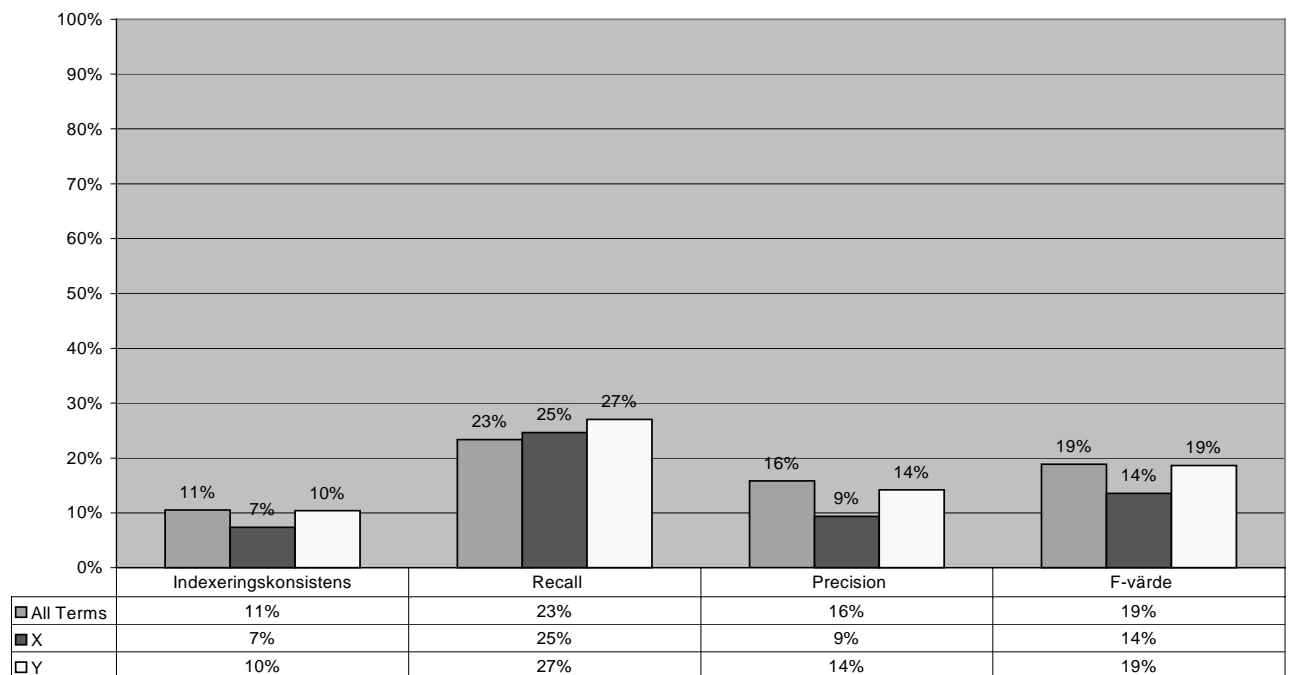


Diagram 33: Följdmotioner, Conexors tio högst rankade termer och indexerarnas termer

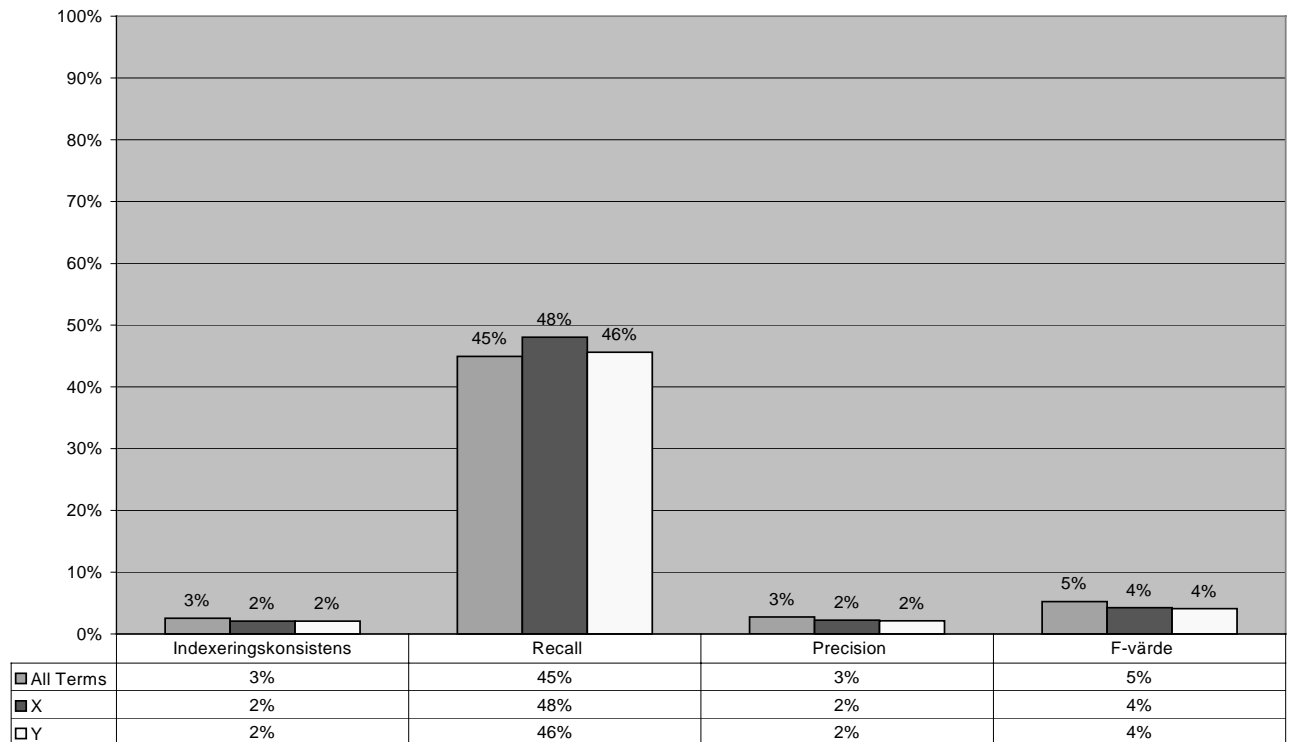


Diagram 34: Interpellationer, Conexors alla termer och indexerarnas termer

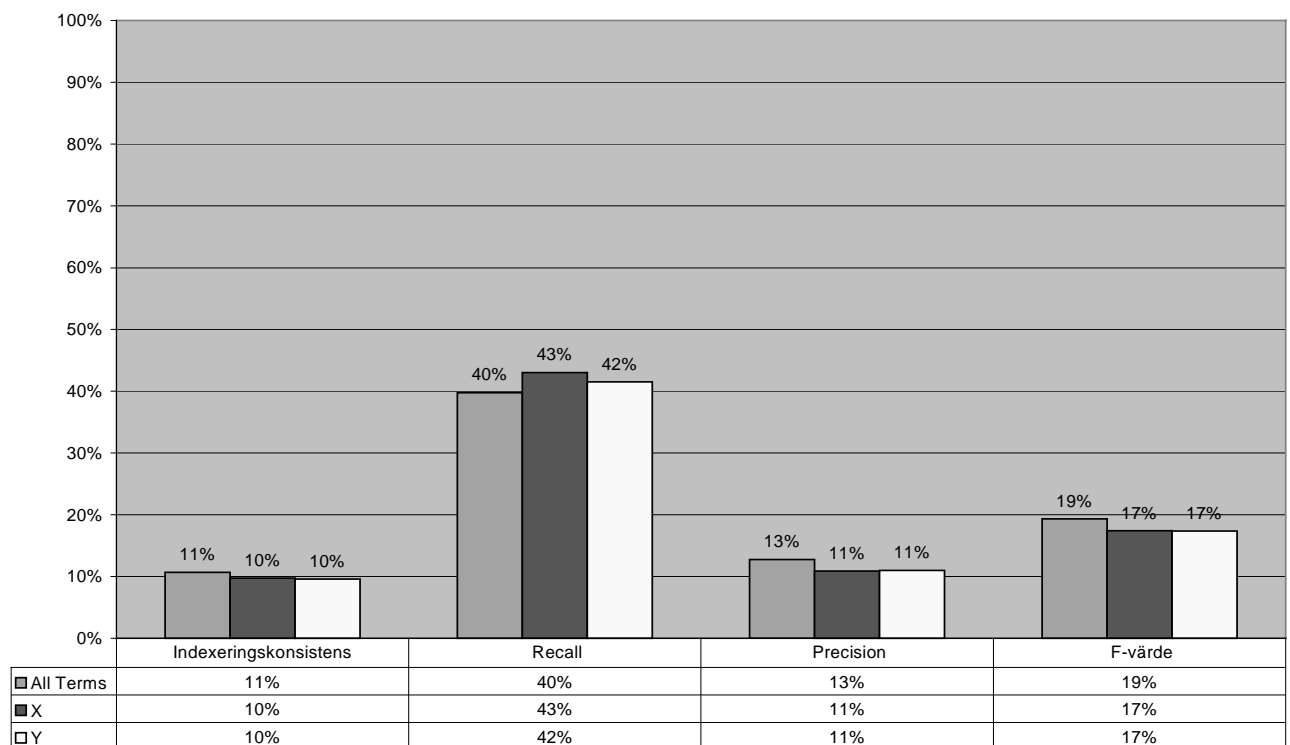


Diagram 35: Interpellationer, Conexors filtrerade termer och indexerarnas termer

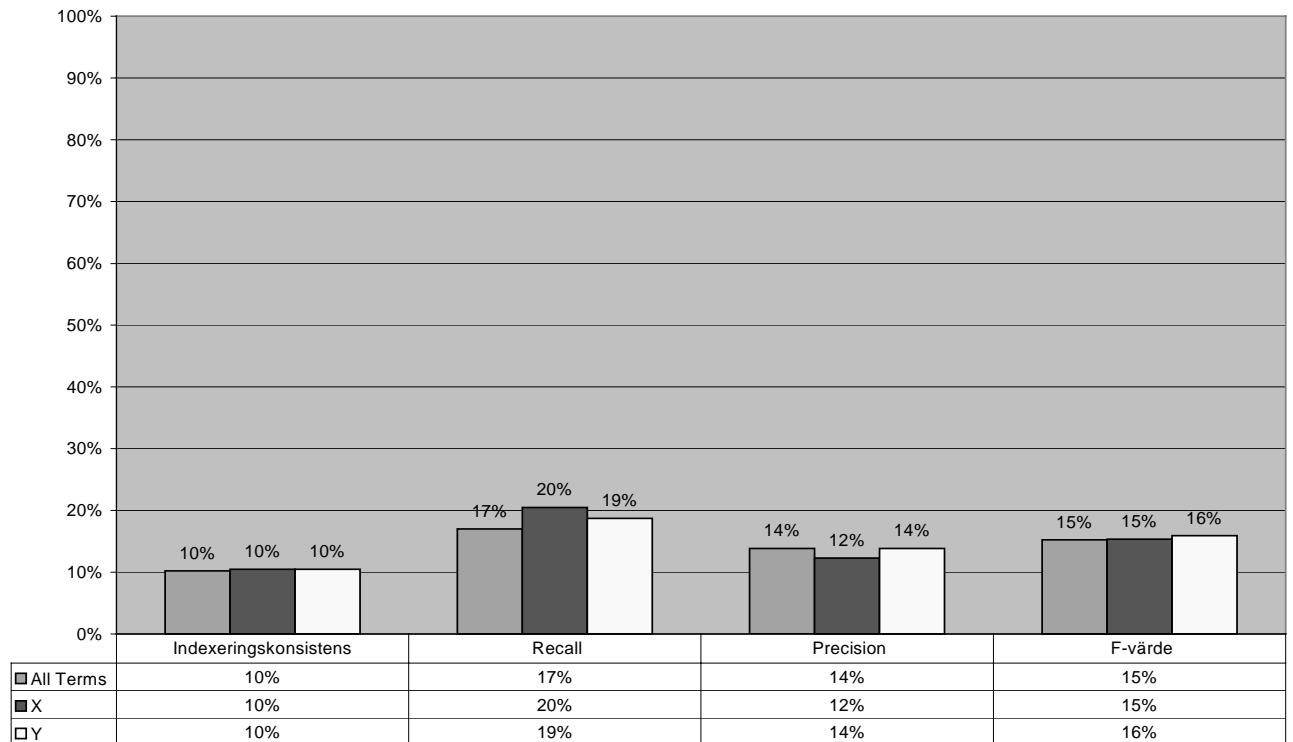


Diagram 36: Interpellationer, Conexors fem högst rankade termer och indexerarnas termer

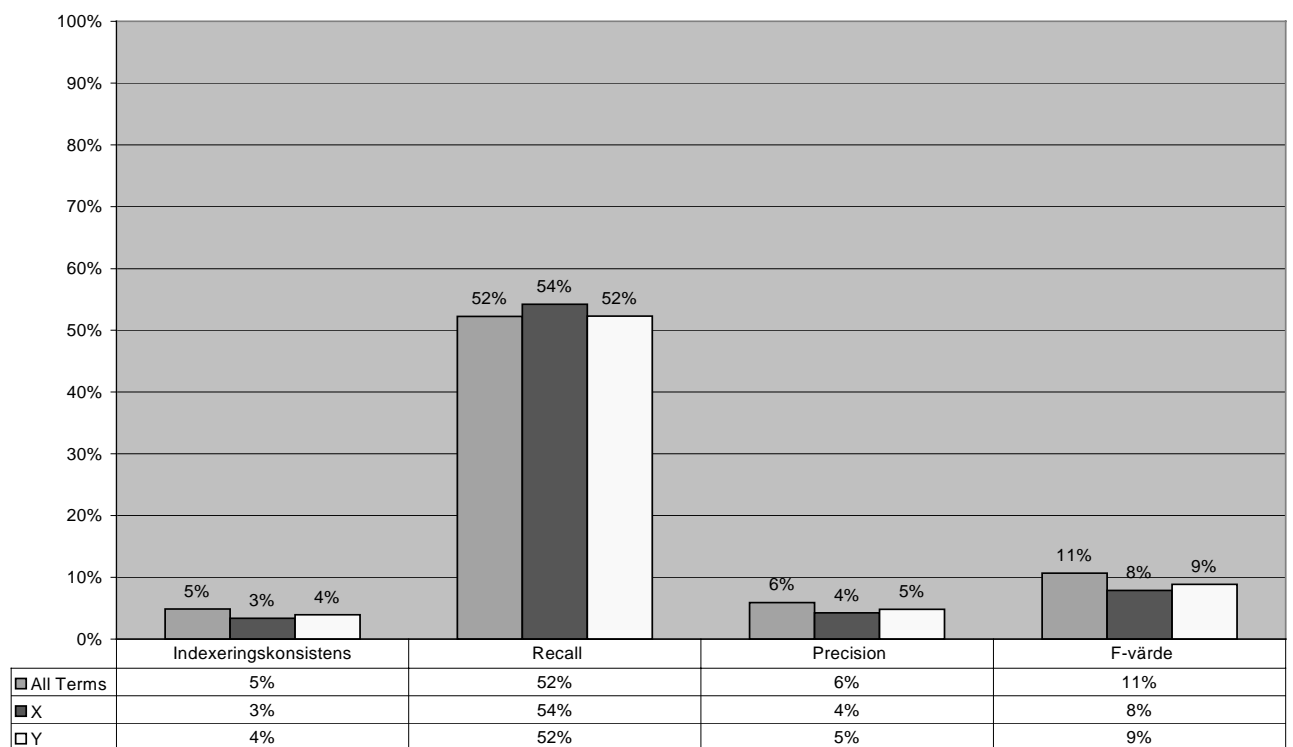


Diagram 37: Frågor, Conexors alla termer och indexerarnas termer

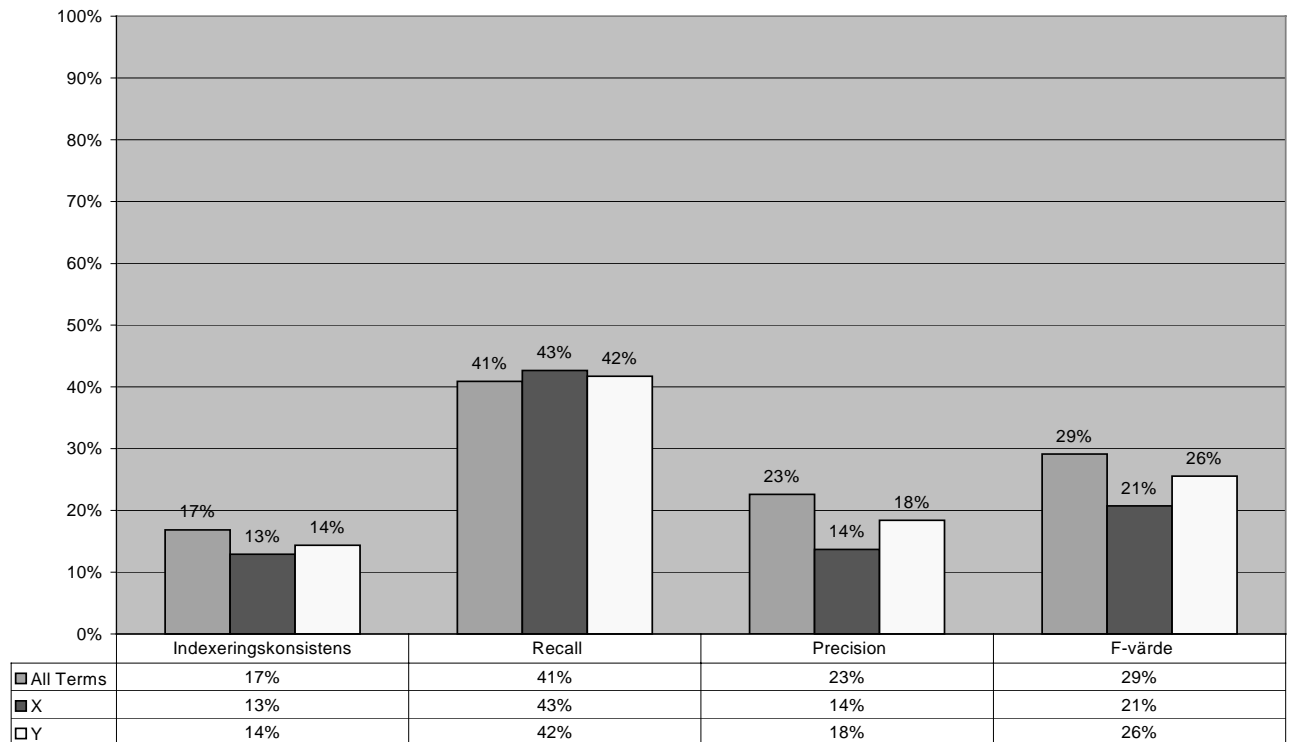


Diagram 38: Frågor, Conexors filtrerade termer och indexerarnas termer

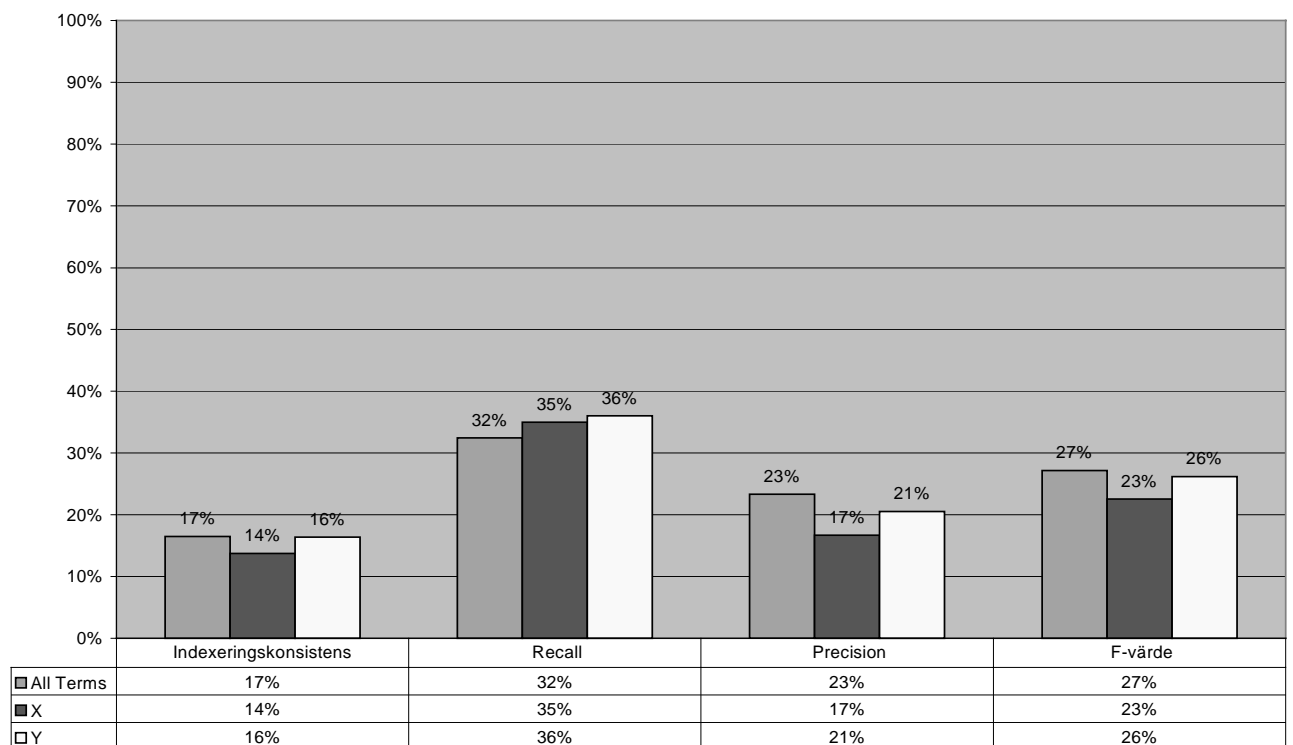


Diagram 39: Frågor, Conexors fem högst rankade termer och indexerarnas termer

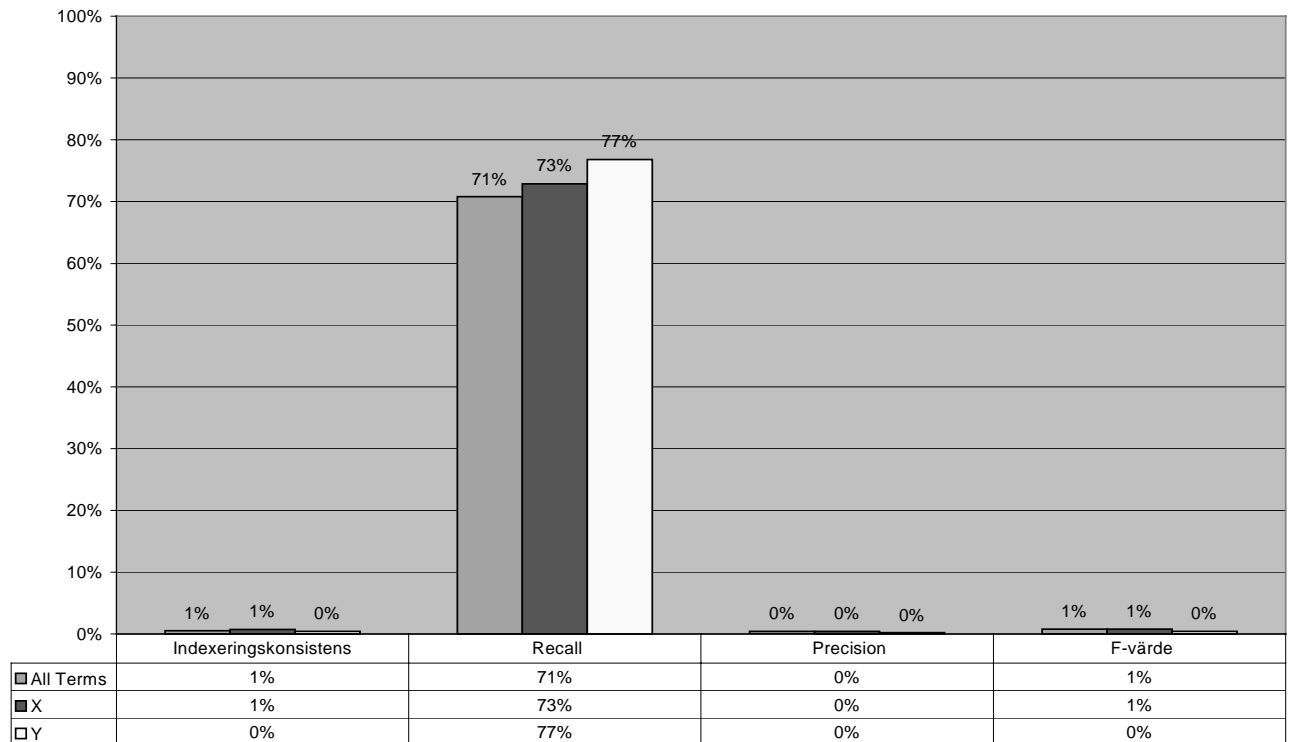


Diagram 40: Propositioner, Conexors alla termer och indexerarnas termer

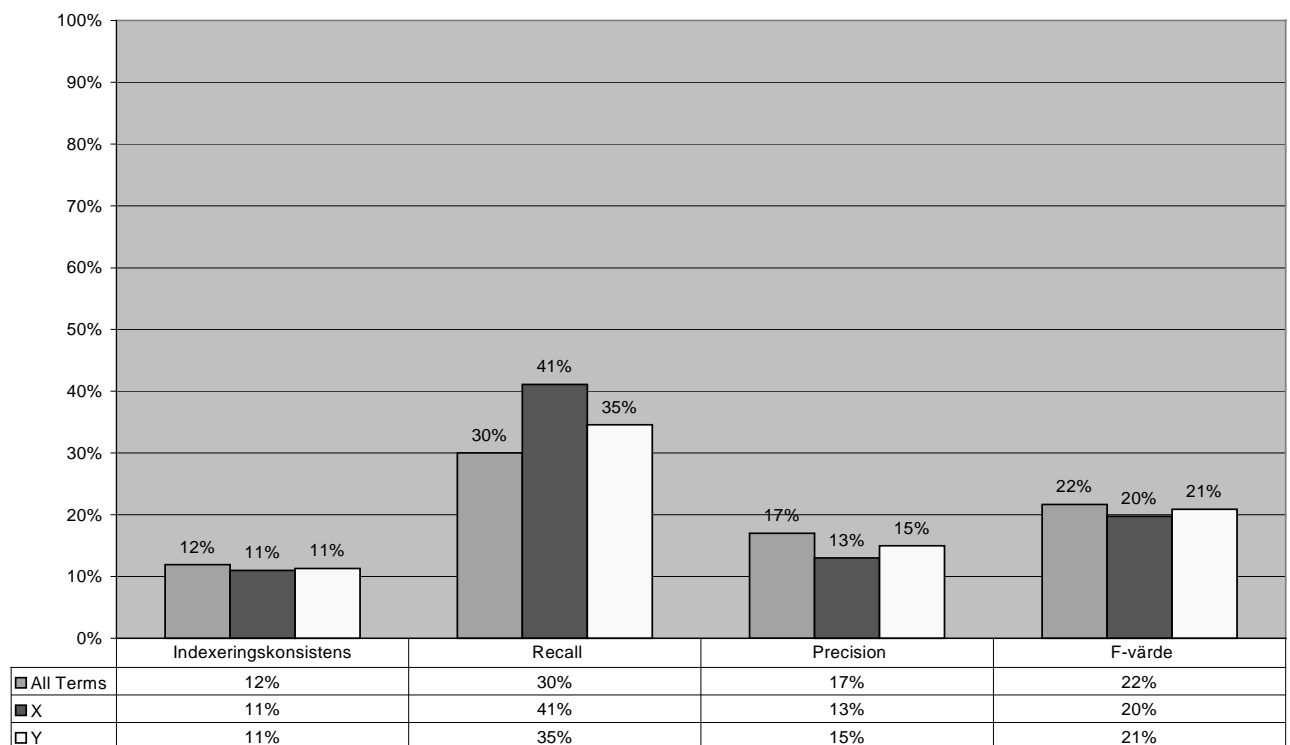


Diagram 41: Propositioner, Conexors tjugo högst rankade termer och indexerarnas termer

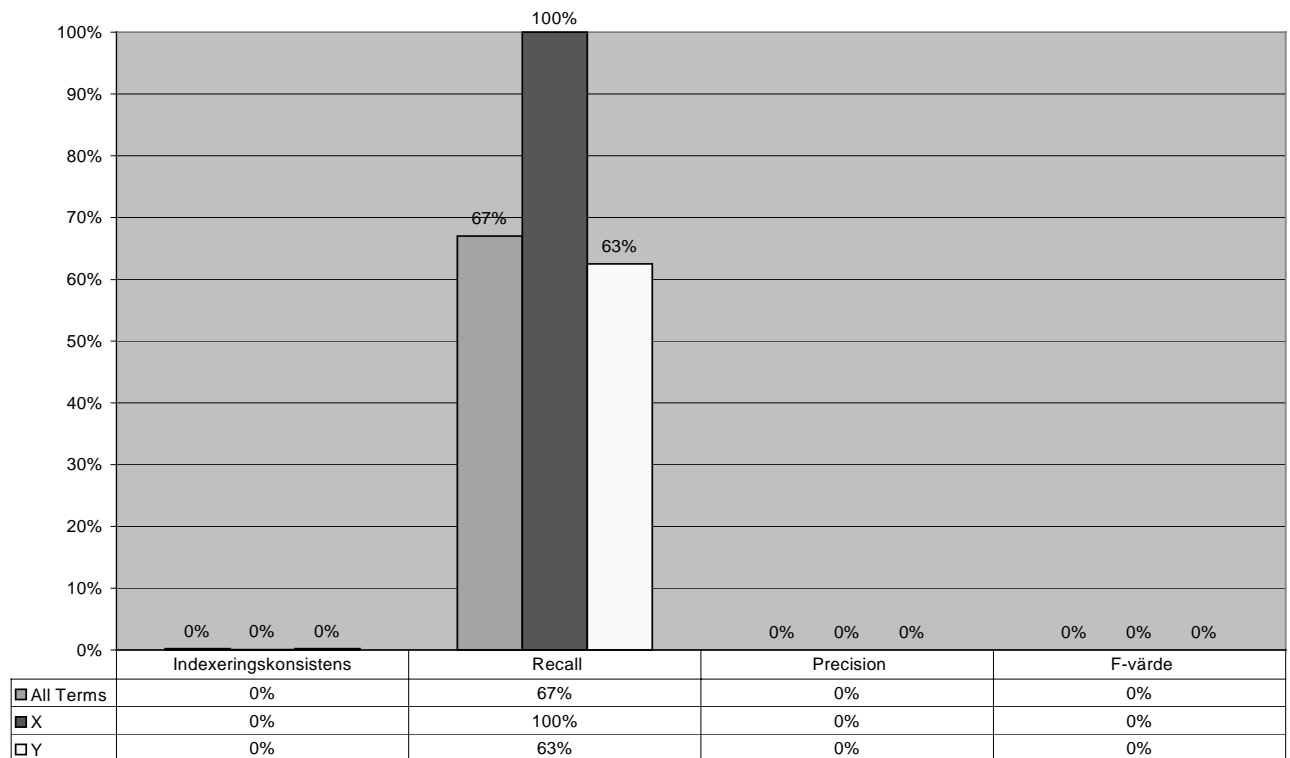


Diagram 42: Skrivelse, Conexors alla termer och indexerarnas termer

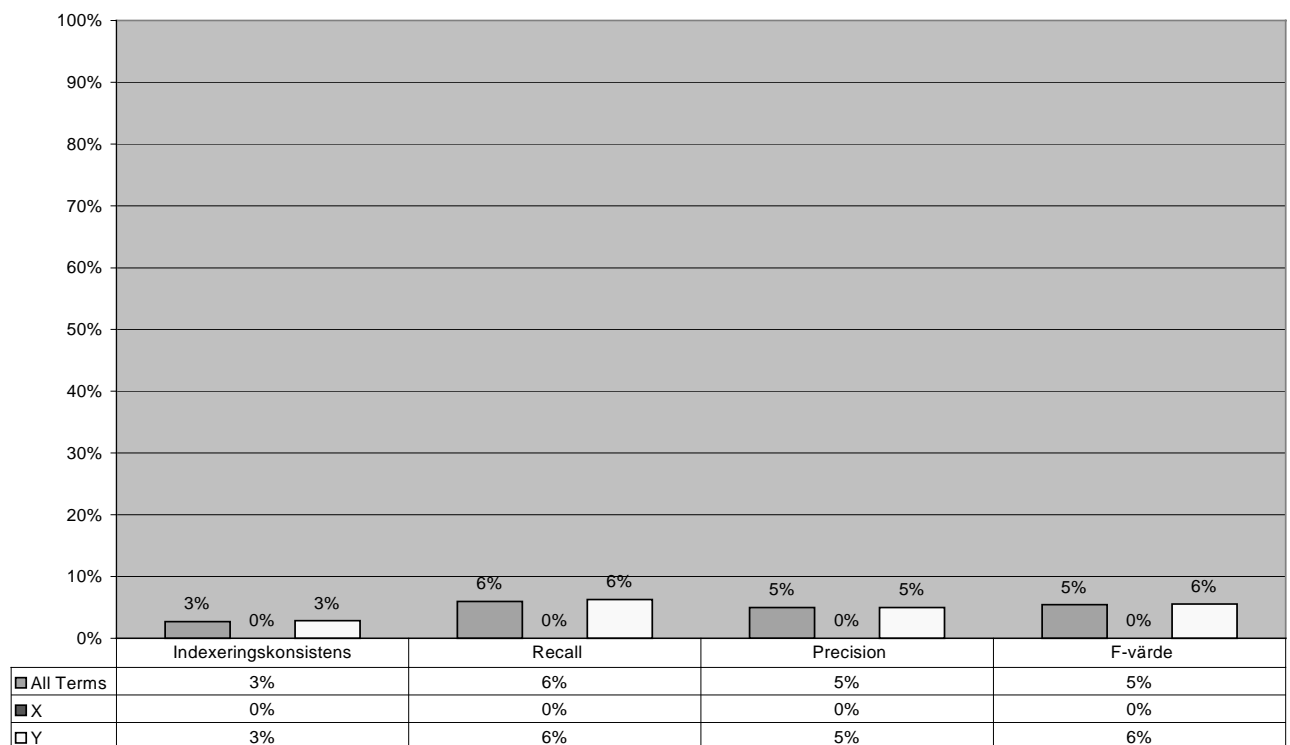


Diagram 43: Skrivelse, Conexors tjugo högst rankade termer och indexerarnas termer

Bilaga LexWare Labs

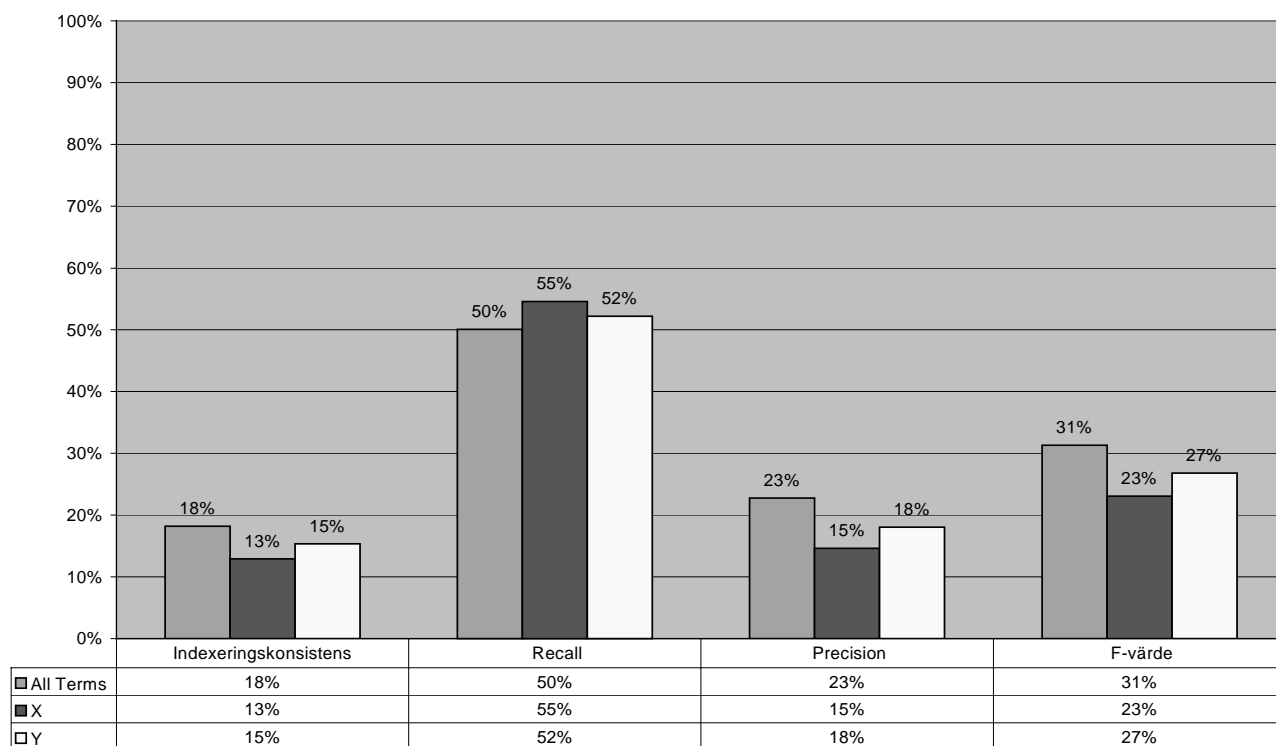


Diagram 44: Allmänna motioner, LexWare Labs alla termer och indexerarnas termer

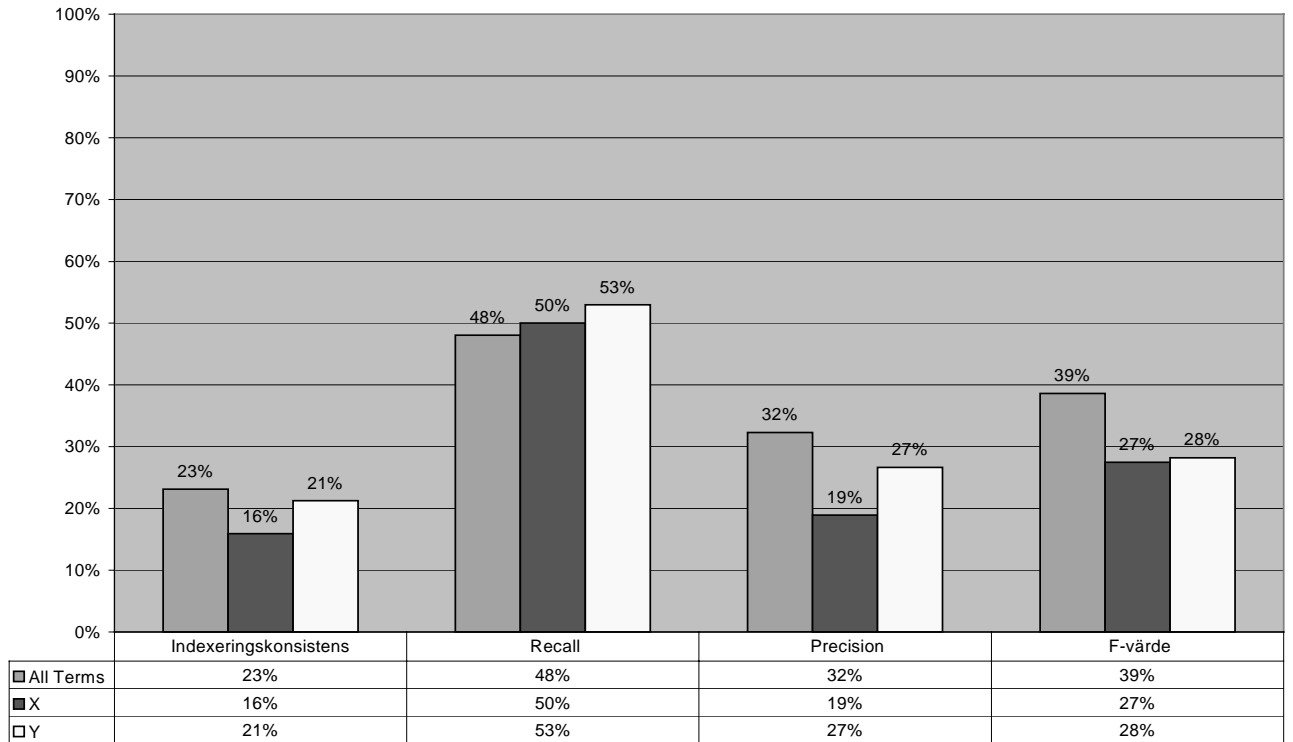


Diagram 45: Allmänna motioner, LexWare Labs tio högst rankade termer och indexerarnas termer

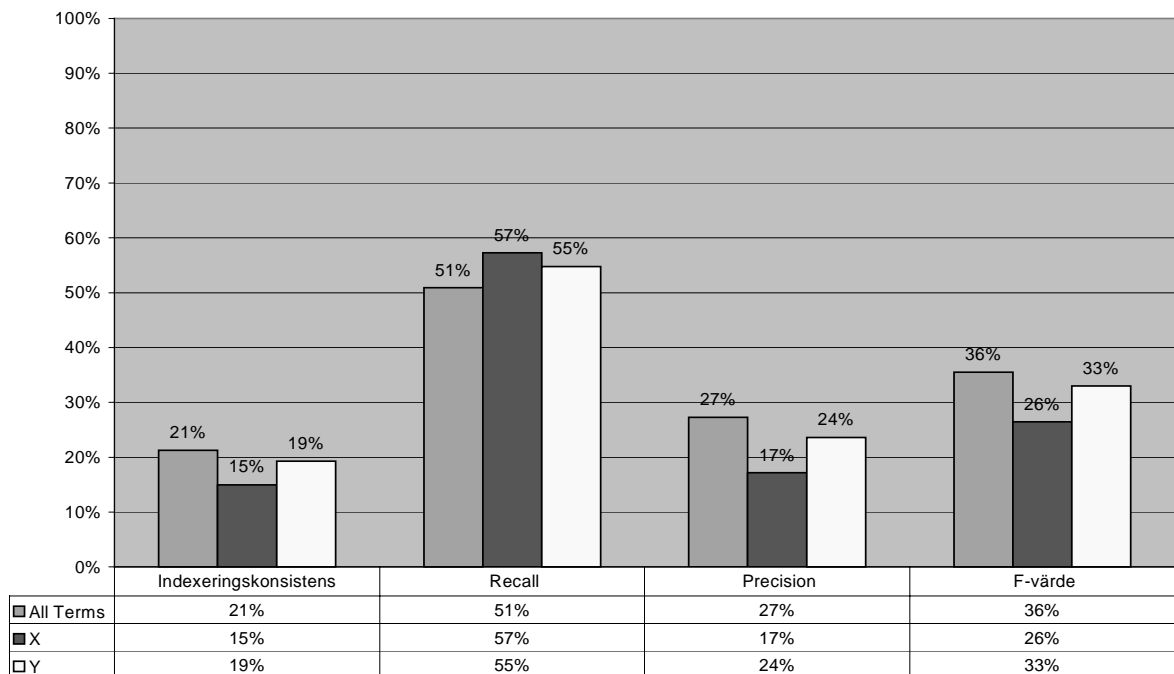


Diagram 46: Följdmotioner, LexWare Labs alla termer och indexerarnas termer

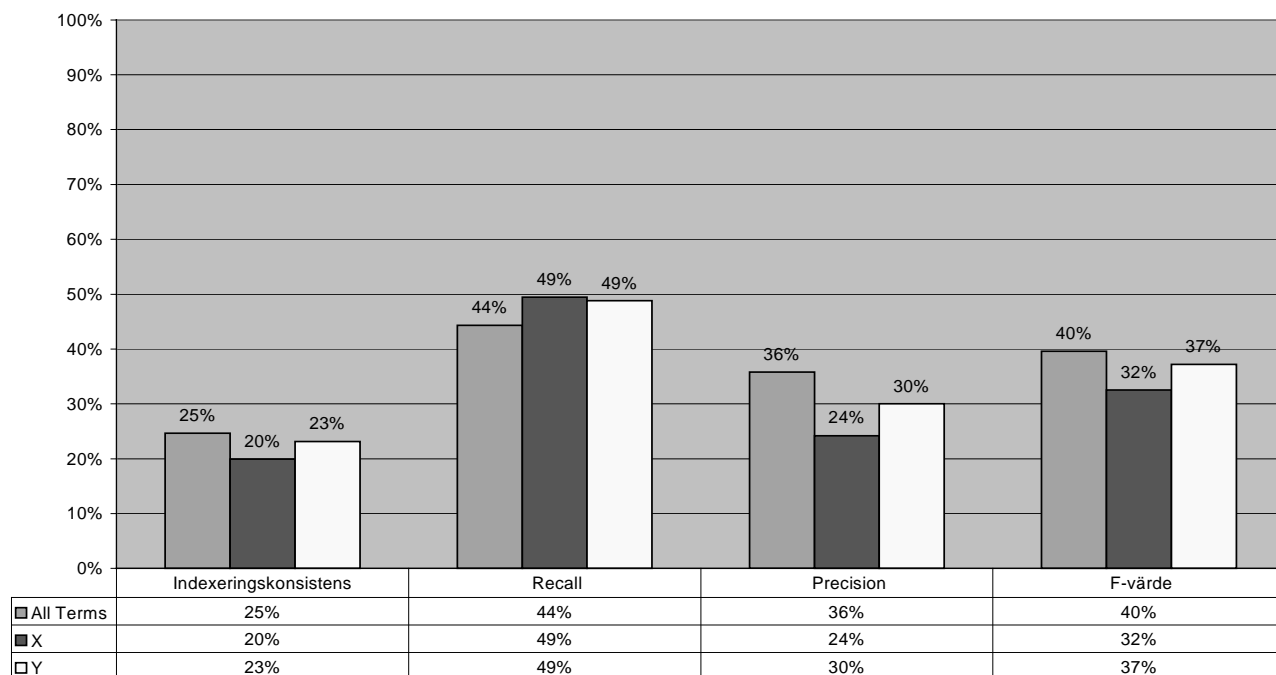


Diagram 47: Följdmotioner, LexWare Labs tio högst rankade termer och indexerarnas termer

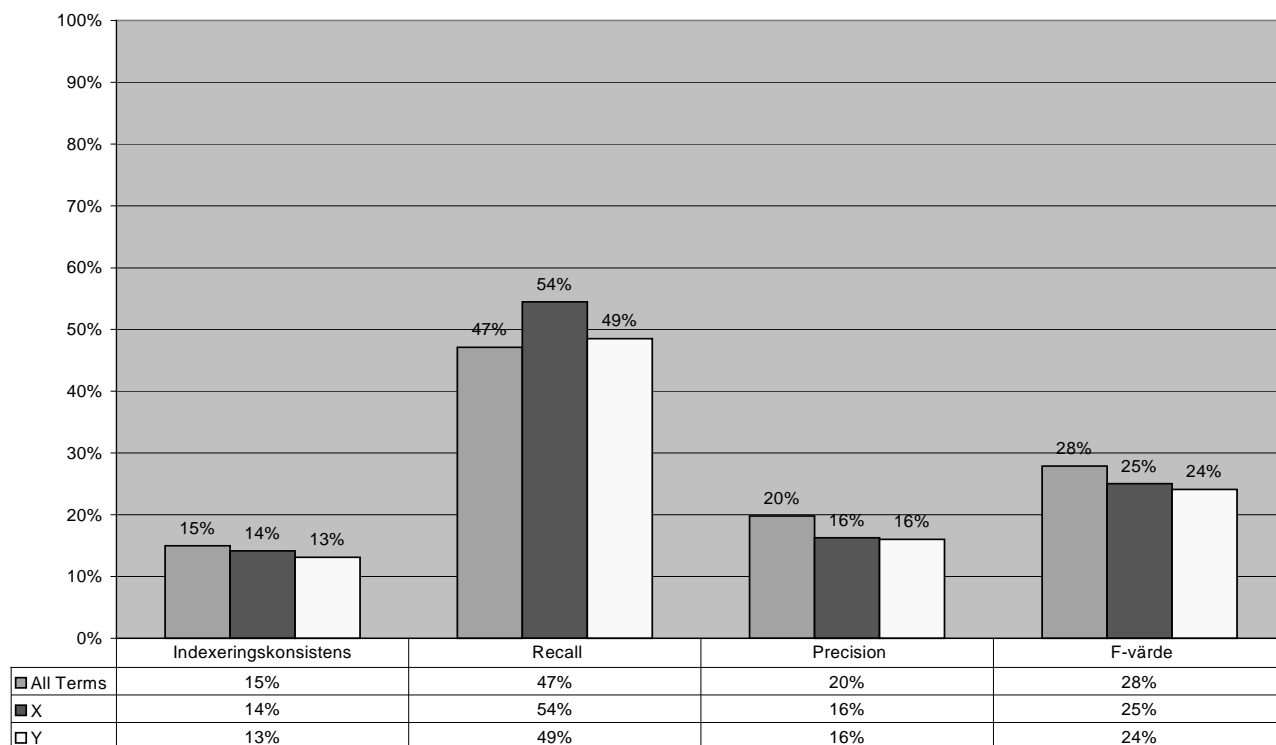


Diagram 48: Interpellationer, LexWare Labs alla termer och indexerarnas termer

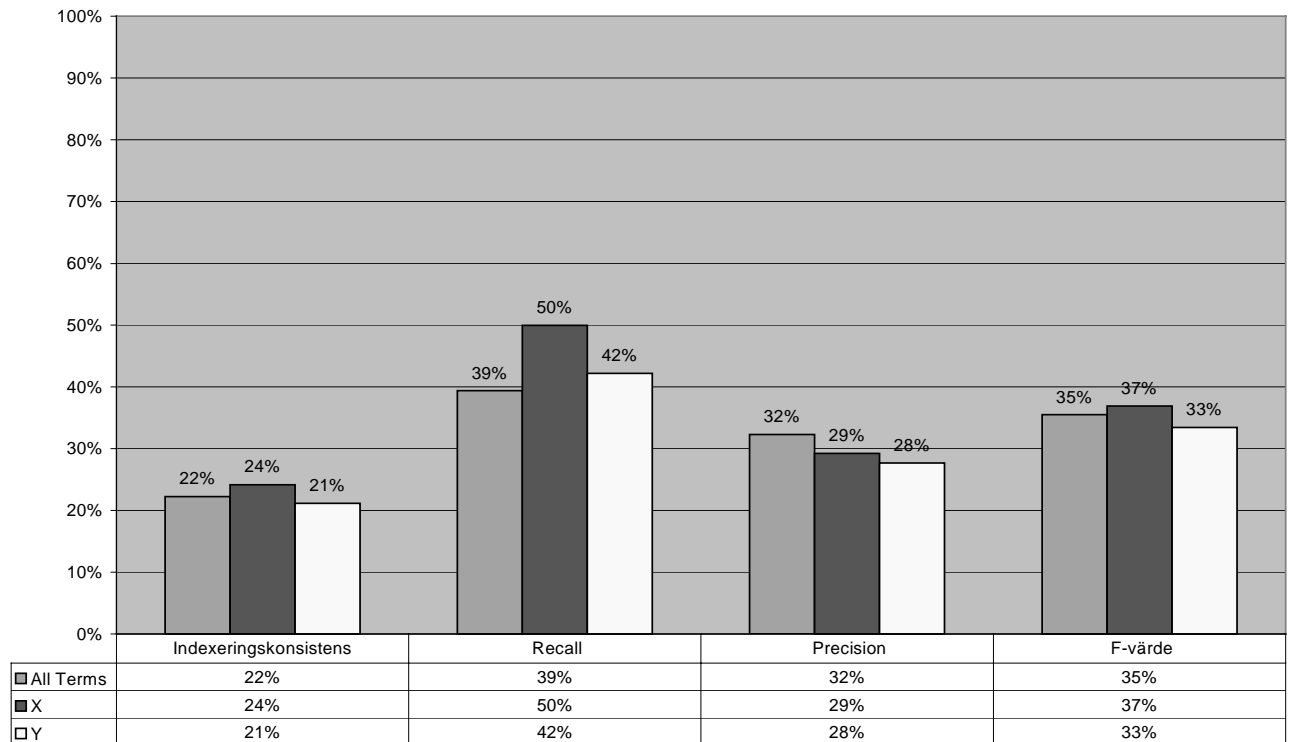


Diagram 49: Interpellationer, LexWare Labs fem högst rankade termer och indexerarnas termer

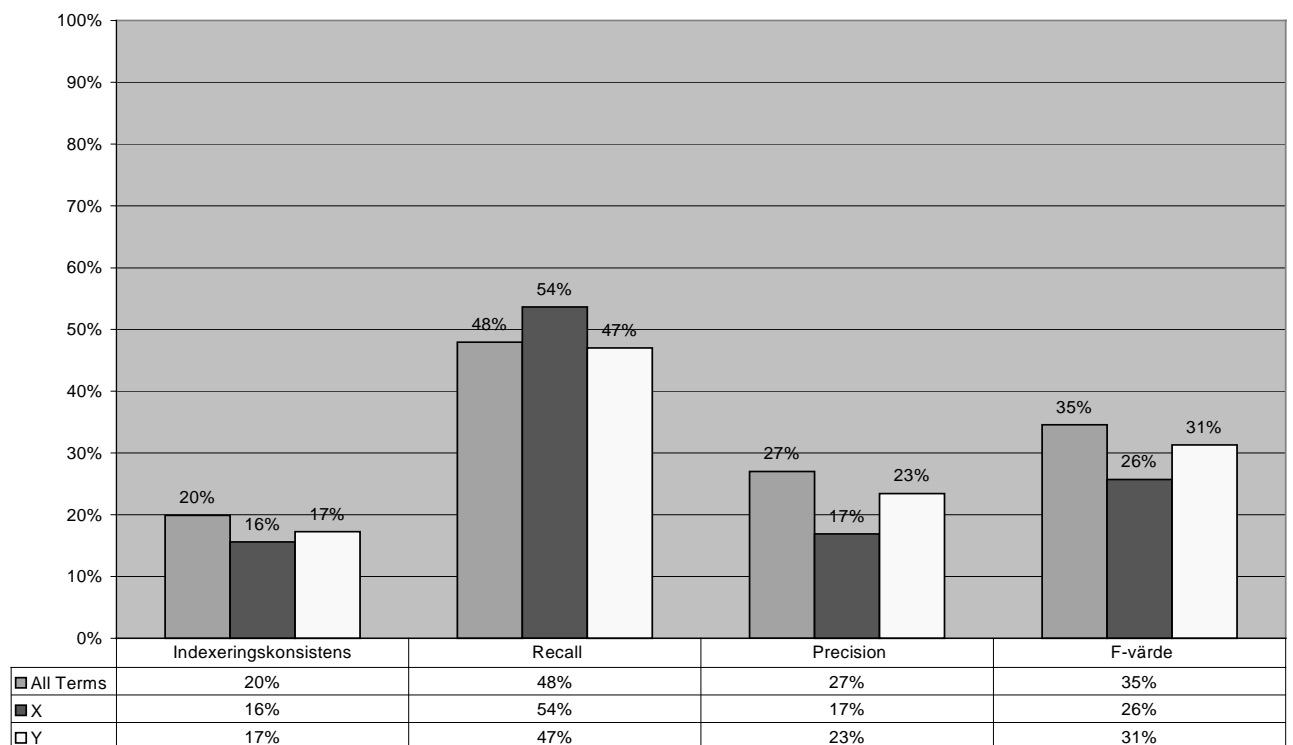


Diagram 50: Frågor, LexWare Labs alla termer och indexerarnas termer

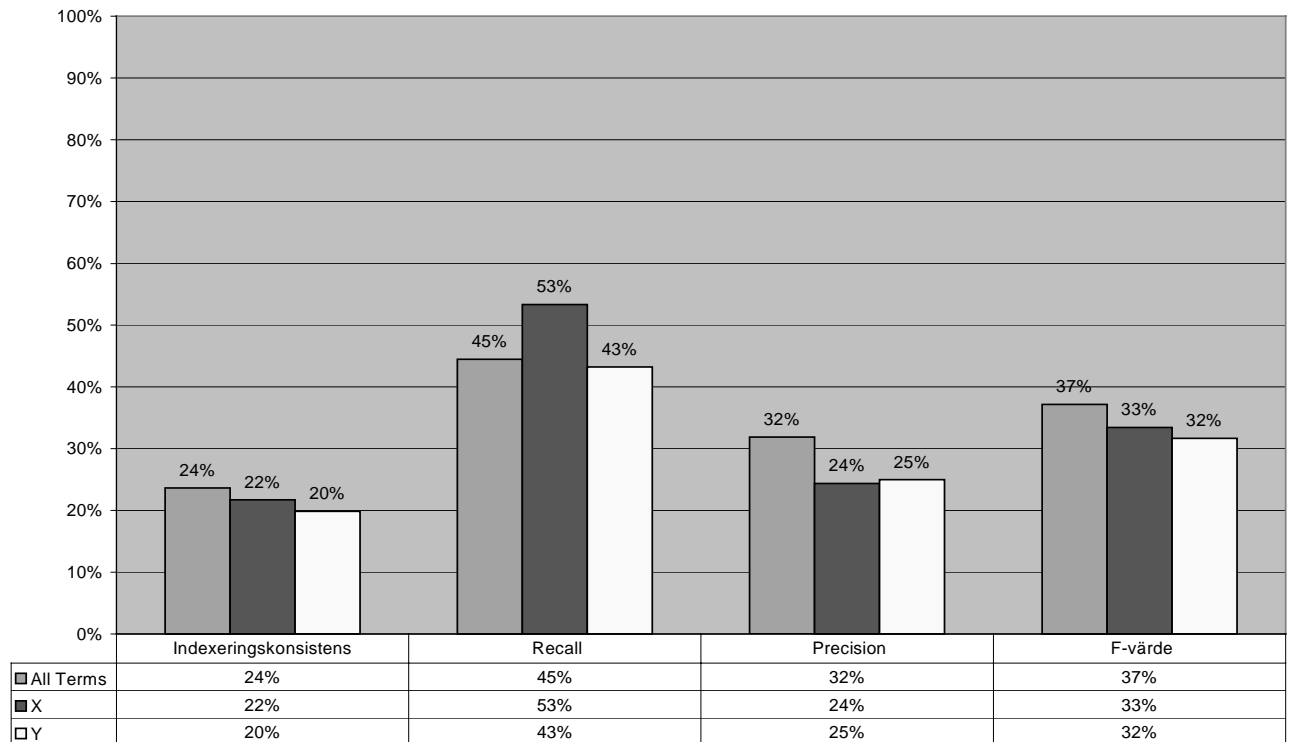


Diagram 51: Frågor, Lex Ware Labs fem högst rankade termer och indexerarnas termer

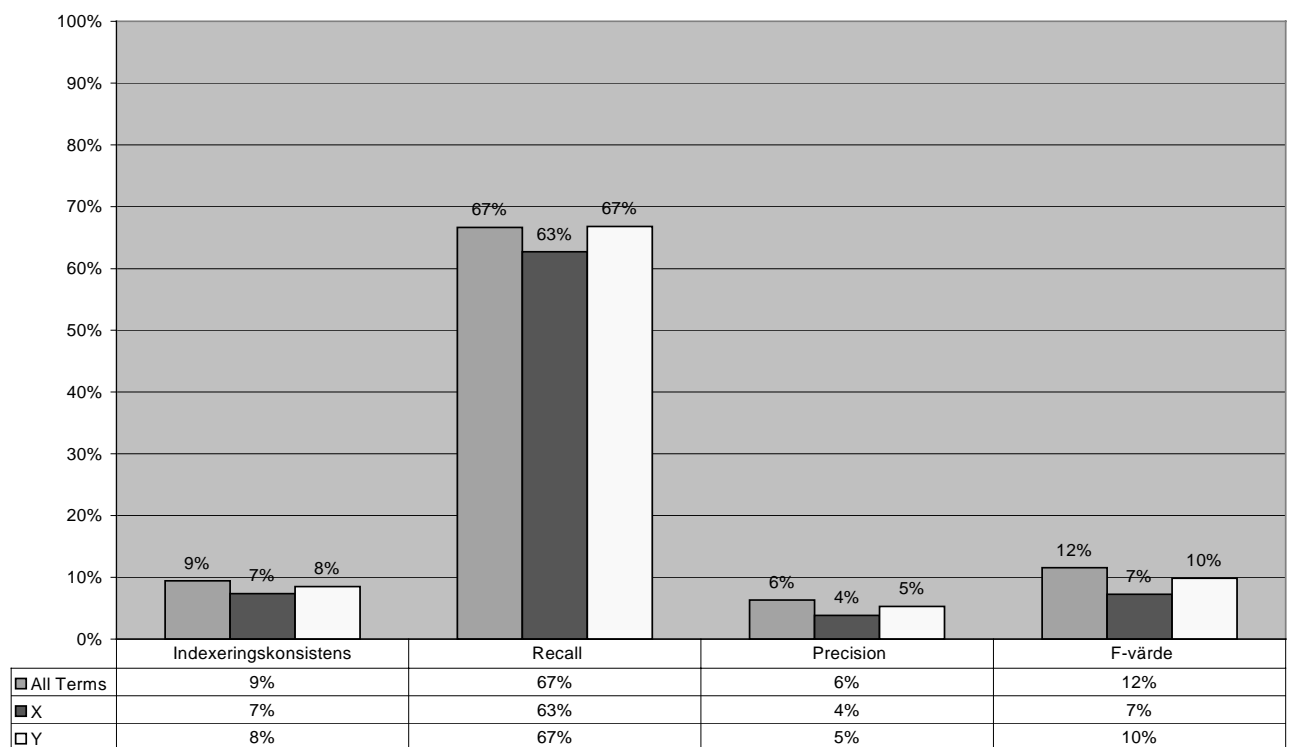


Diagram 52: Propositioner, LexWare Labs alla termer och indexerarnas termer

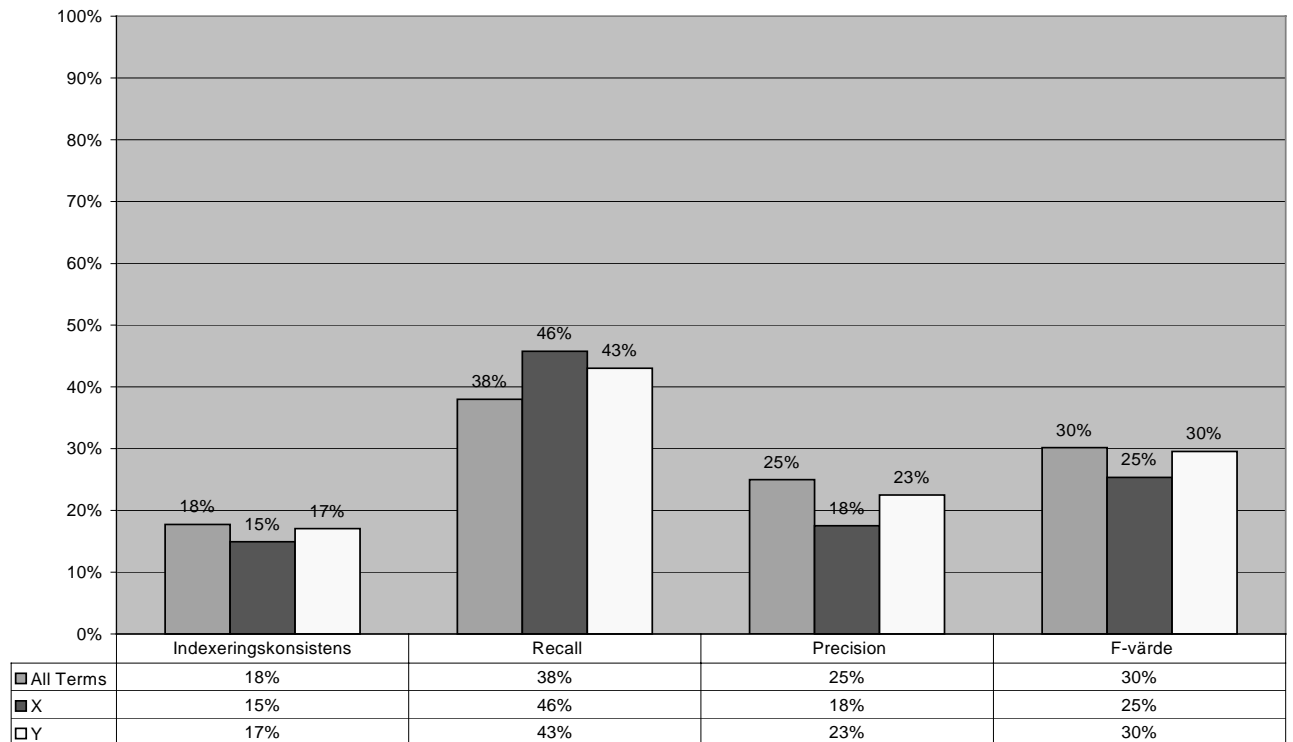


Diagram 53: Propositioner, LexWare Labs tjugo högst rankade termer och indexerarnas termer

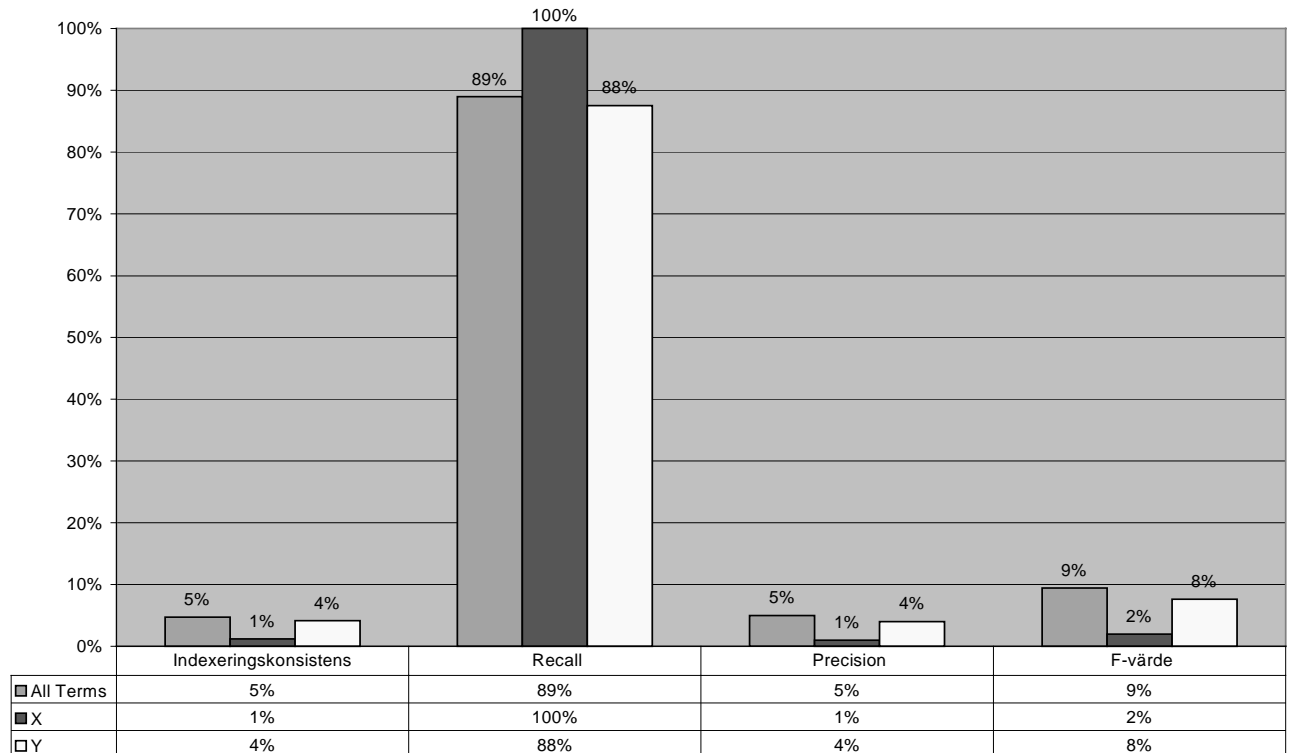


Diagram 54: Skrivelse, LexWare Labs alla termer och indexerarnas termer

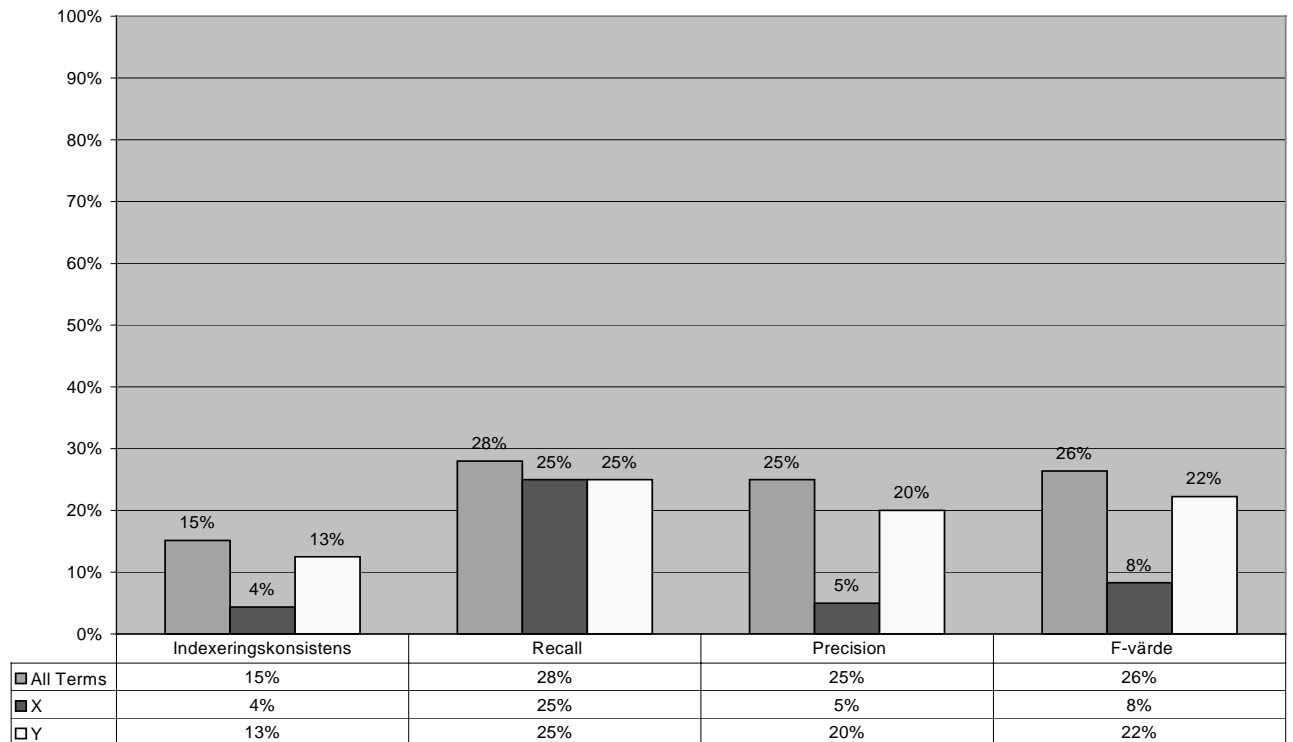


Diagram 55: Skrivelse, LexWare Labs tjugo högst rankade termer och indexerarnas termer

Bilaga KTH

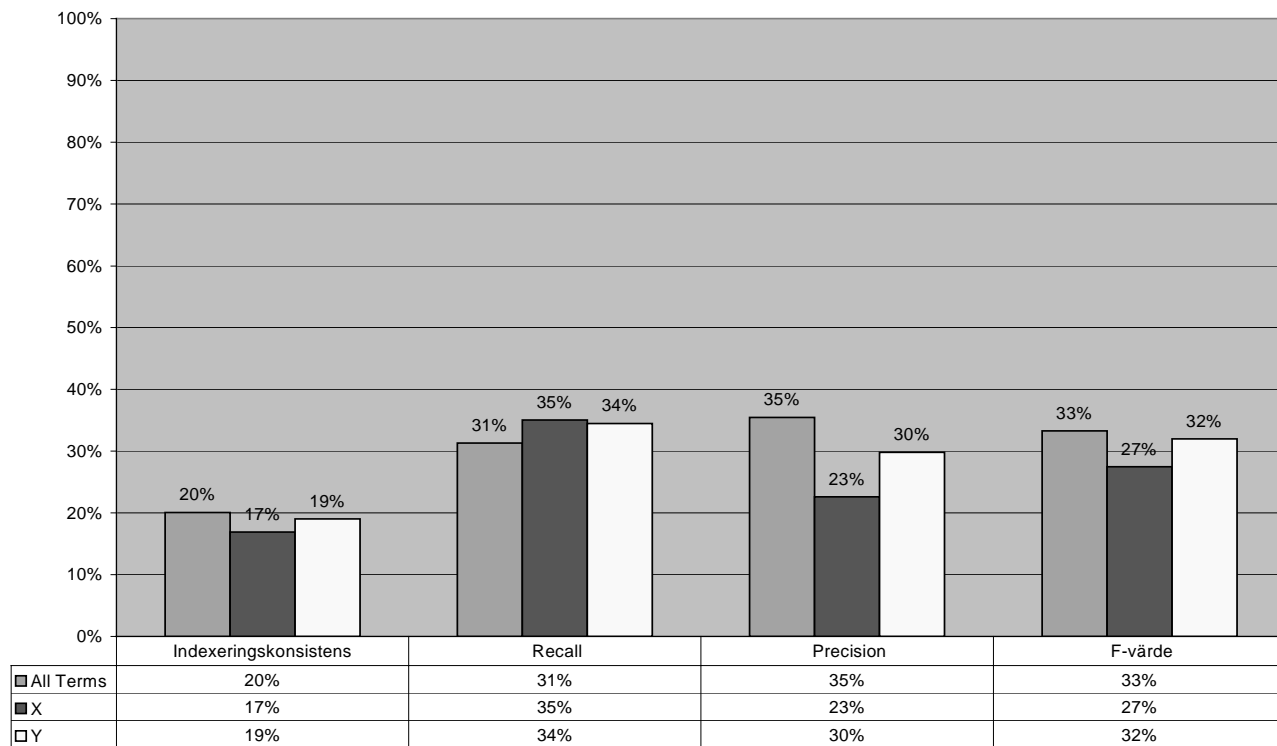


Diagram 56: Allmänna motioner, KTH:s termer och indexerarnas termer

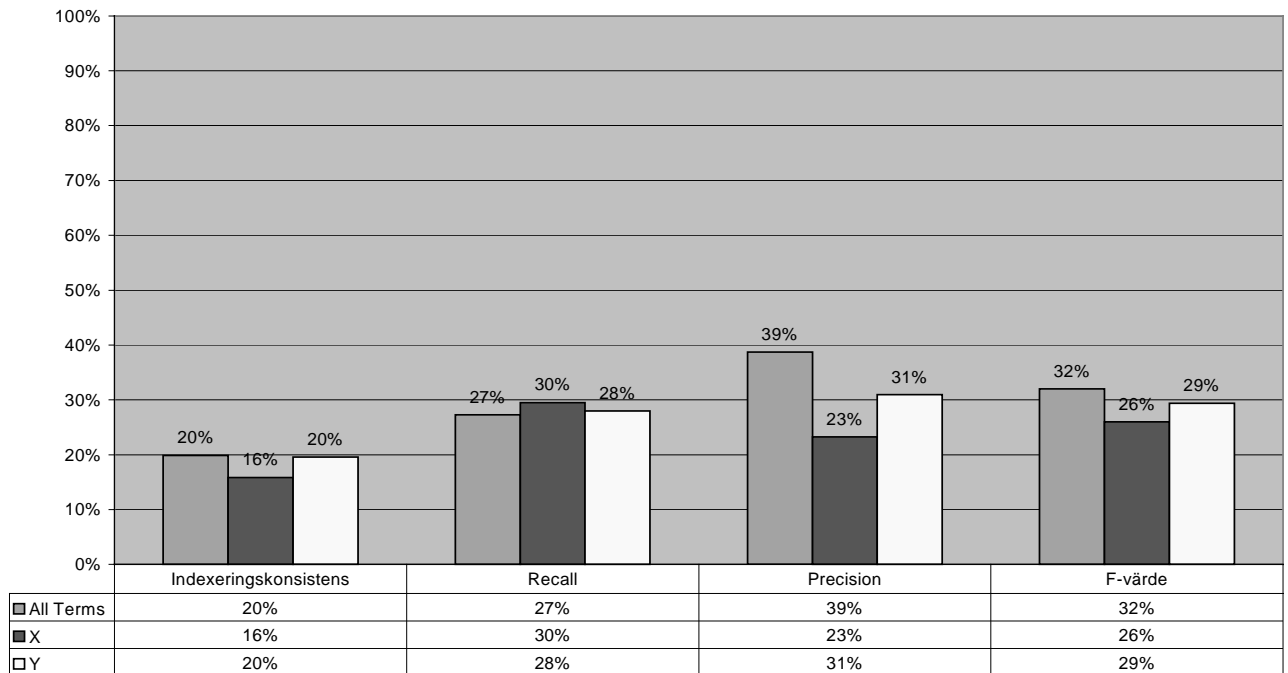


Diagram 57: Följdmotioner, KTH:s termer och indexerarnas termer

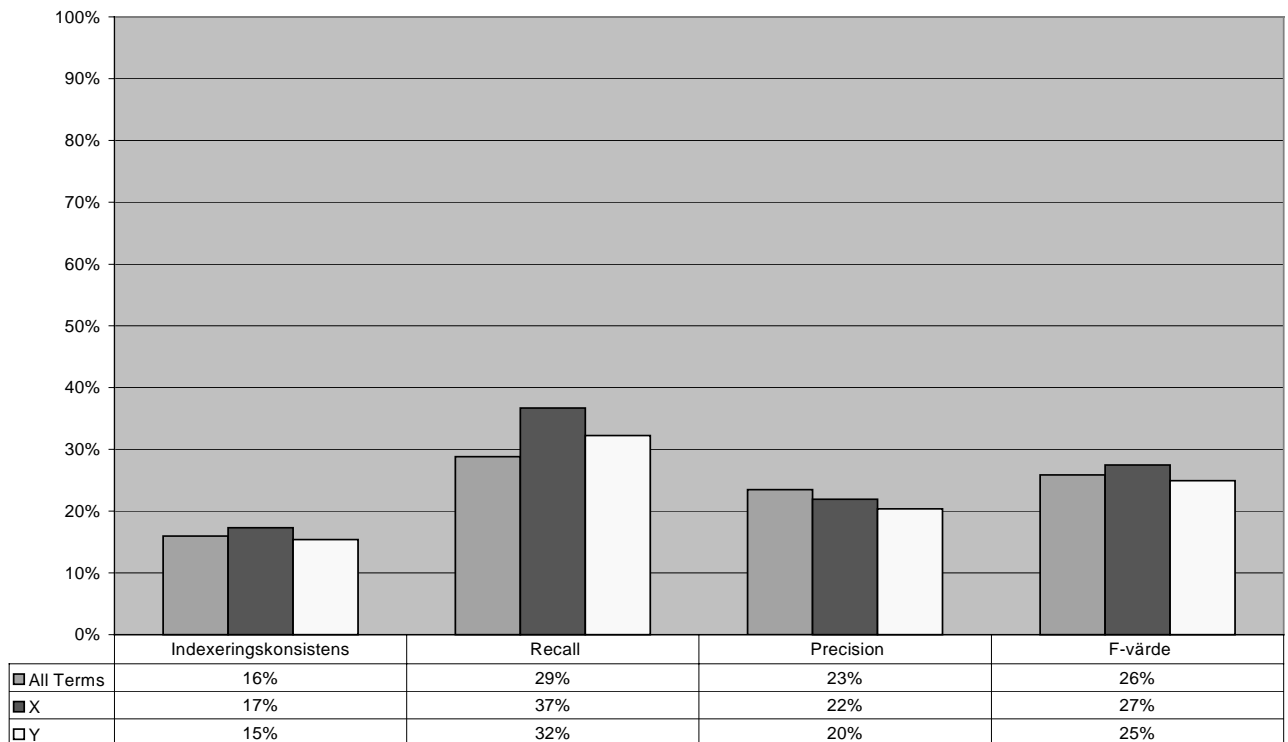


Diagram 58: Interpellationer, KTH:s termer och indexerarnas termer

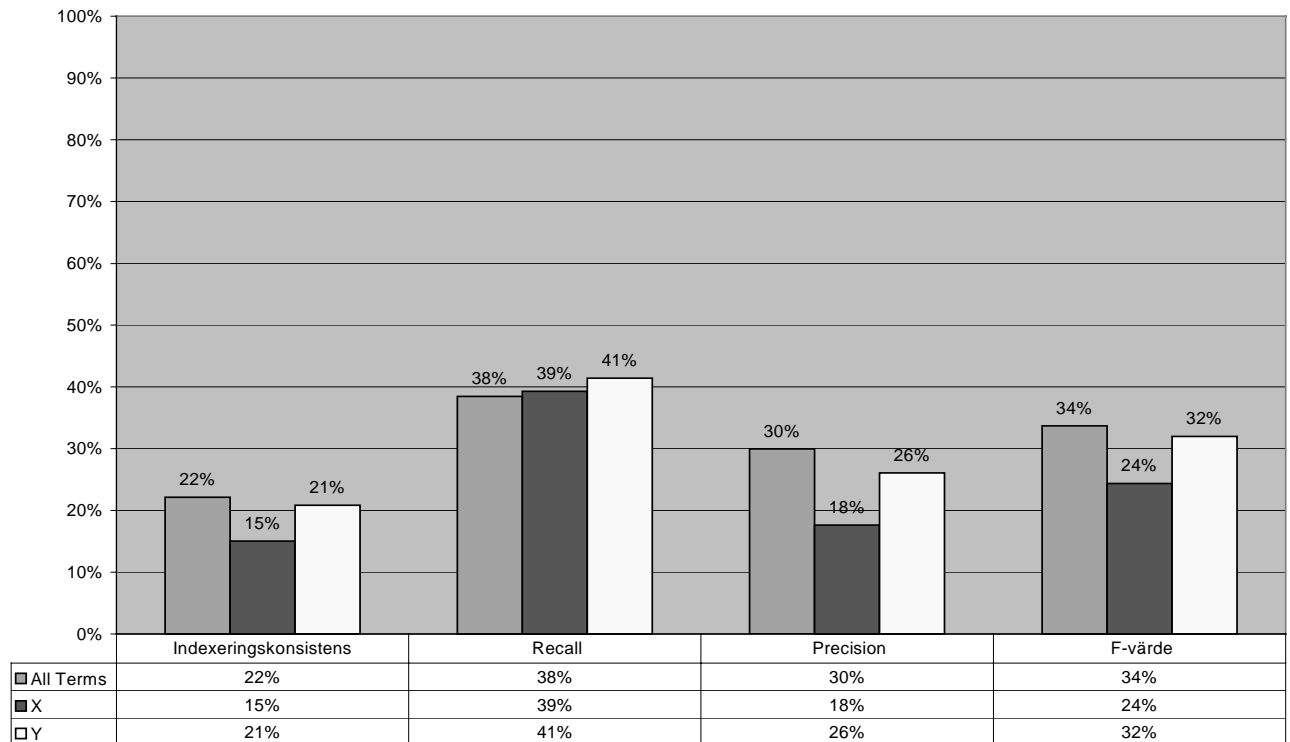


Diagram 59: Frågor, KTH:s termer och indexerarnas termer

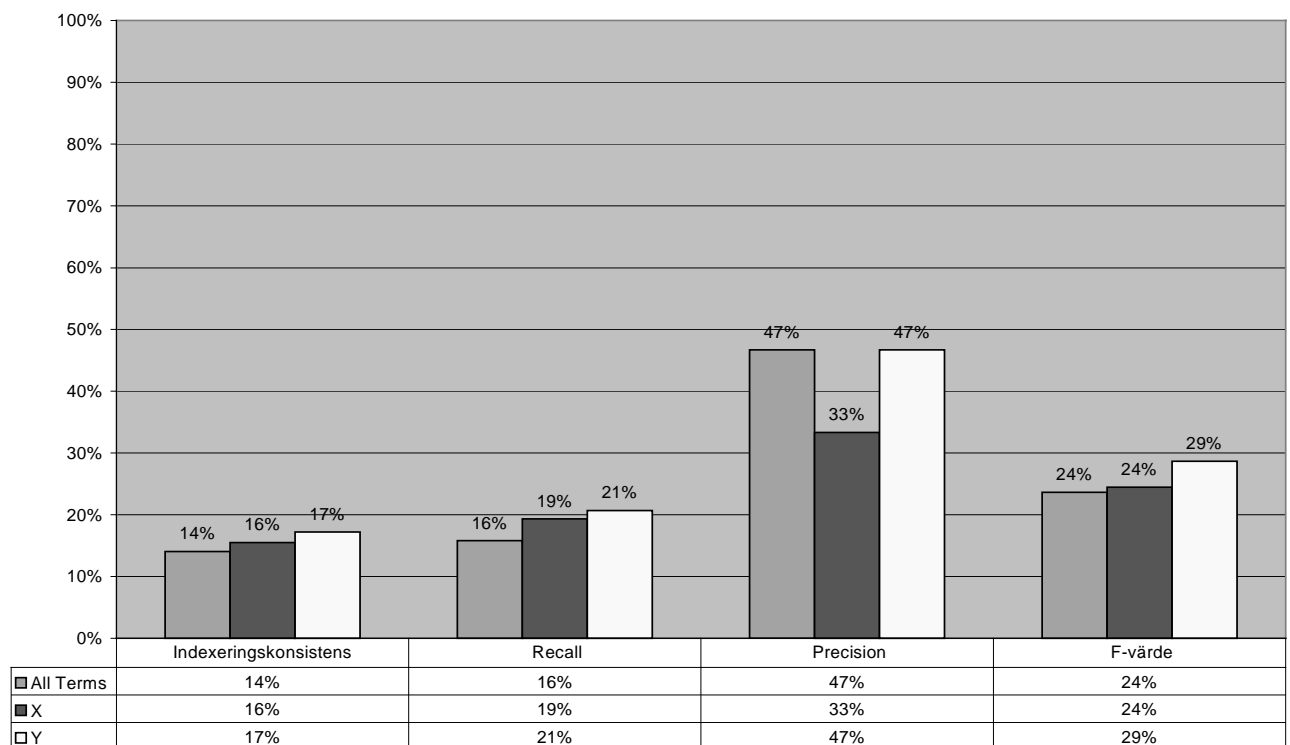


Diagram 60: Propositioner, KTH:s termer och indexernas termer

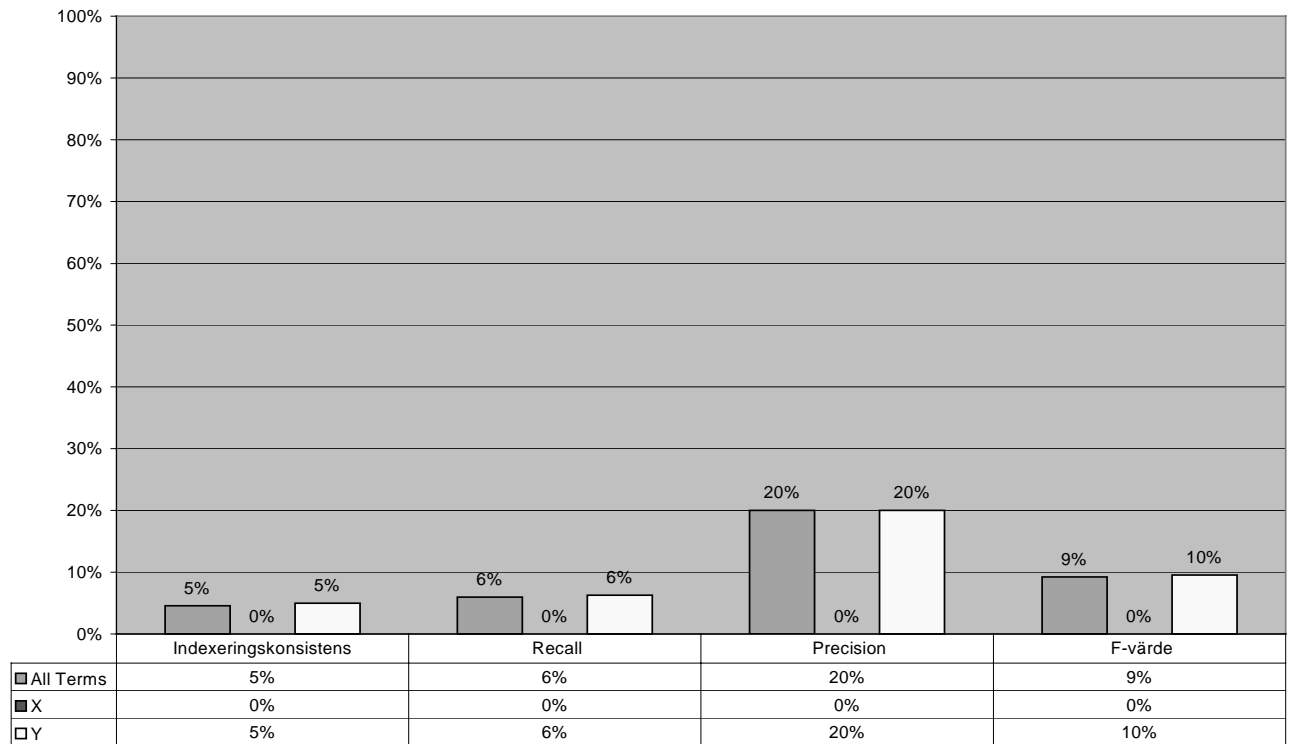


Diagram 61: Skrivelse, KTH:s termer och indexerarnas termer