MAGISTERUPPSATS I BIBLIOTEKS- OCH INFORMATIONSVETENSKAP
VID BIBLIOTEKS- OCH INFORMATIONSVETENSKAP/BIBLIOTEKSHÖGSKOLAN
2005:121
ISSN 1404-0891

Lexikonbaserad Cross-Language Information Retrival:
Utvärdering av queryeffektivitet

FANGLAN CHEN
TAISSIA GORIOUNOVA

HÖGSKOLAN I BORÅS

Svensk titel: Lexikonbaserad Cross-Language Information Retrival: Utvärdering av queryeffektivitet

Engelsk titel: The Dictionary-Based Cross-Language Information Retrieval: Evaluation of Query Effectiveness

Författare: Fanglan Chen
Taissia Goriounova

Kollegium: 2

Färdigställt: 2005

Handledare: Per Ahlgren

Abstract: This thesis discusses main problems associated with dictionary-based Cross-Language Information Retrieval as lexical and translational ambiguity of query terms, translation of compounds and phrases, dictionary limitation. The purpose of the study is to investigate how query structure influences the effectiveness of CLIR regarding performance of three query types: original query, unstructured query and structured query. Query structuring refers to the application of #syn-operator to group query terms. The study comprises an experiment that was performed in the InQuery IR system with TrecUta database that contains 550,000 news articles from different American newspapers. 24 topics were used for the experiment. The effectiveness of three types of query structure is compared at different Document Cut-off Value levels, maximal DCV= 100. The measure used is average precision. Binary relevance situation, where the three relevance degrees (1, 2, and 3) have been merged into one, is applied. The results show that dictionary-based query translation without the use of structure significantly decreases the effectiveness of information retrieval while query structuring through synonym sets shows to be a simple and effective method, which allows the reduction of the effects of translation ambiguity and the improvement of the performance of CLIR-queries. The results reveal that the performance can nearly reach the same level as the original queries.

Nyckelord: Cross-Language Information Retrieval, CLIR, queryöversättning, querystruktur, effektivitetsutvärdering

Innehållsförteckning

Förkortningar

CLIR – Cross-Language Information Retrieval

CLEF – Cross-Language Evaluation Forum

TREC – Text Retrieval Conference

NTCIR – NII-NACSIS (National Center for Science Information Systems) Test Collection for IR Systems

MT – Machine Translation

MRD – Machine-Readable Dictionary

DCV– Document Cut-off value

1. Inledning

Cross-Language Information Retrieval (CLIR) är ett relativt nytt forskningsområde inom information retrieval (IR) som hanterar informationsåtervinning ur ett flerspråkigt perspektiv. CLIR innebär huvudsakligen en möjlighet att formulera query (en söksträng) på ett annat språk än dokumentspråket i textkollektion och återvinna dokument på olika språk.

CLIR är en nödvändighet idag. Den följer med utvecklingen av Internet och relaterad teknik, samt med växande informationsflöde på olika språk (Pirkola, 1998; Oard & Dorr, 1996). Trots att engelska och andra språk som japanska, tyska, franska o.s.v., som talas i högt ekonomiskt utvecklade länder, fortfarande är dominerande språk på webben kommer andra språk som kinesiska, spanska, arabiska och många andra som talas av en stor folkmängd att få större vikt.

För att möjliggöra cross-language information retrieval brukar man översätta antingen dokument eller query. Det reducerar problemet till enspråkig IR. Både dokument- och queryöversättning har sina fördelar och nackdelar. Dokumentöversättning kan nå högre grad av noggrannhet p.g.a. att hela dokumentet tillför en rikare kontext. Men den kräver större resurser och kostar mer eftersom hela dokumentsamlingen måste översättas. Numera koncentrerar man sig mest på queryöversättning eftersom det är billigare och mindre tids- och resurskrävande jämfört med översättning av hela dokument. Dessutom skiljer man på olika metoder för queryöversättning, bland annat maskinöversättning, lexikonbaserad/tesaurusbaserad och korpusbaserad översättning.

Lexikonbaserad queryöversättning är en enklare metod jämfört med de andra metoderna. För den används tvåspråkiga lexikon (ofta maskinläsbara) som finns på de flesta spåken. Metoden baseras på att varje ord i källspråket ersätts med alla översättningsekvivalent i målspråket vilka tas till slutlig queryformulering. De grundläggande problem som förknippas även med den lexikonbaserade queryöversättningsmetoden är bl.a. frasöversättning, översättningsflertydighet och problem som orsakas av lexikonets omfång.

Denna studie är ägnad åt tvåspråkig informationssökning med naturligt språk (i kontrast till kontrollerad vokabulär) där engelska är målspråk och svenska är källspråk. Vi kommer att koncentrera oss på lexikonbaserad queryöversättning. Fokus liggar på hur man kan övervinna problem som översättningen medför med hjälp av querystruktur. Querystruktur innebär den syntaktiska strukturen av en query, vilken uttrycks genom olika operatorer och parentes. Vi ska undersöka återvinningseffektivitet av tre querytyper, en engelsk originalquery och två CLIR-queries: den ena ostrukturerad och den andra strukturerad. En ostrukturerad query innebär en mekaniskt ihopsättning av alla översättningar av alla termer. En strukturerad query innebär att alla lexikonets synonymer grupperas i samma facett med hjälp av *synonymi*-operator. Undersökningen ska genomföras i experimentell miljö, och ett probabilistiskt IR-system, InQuery, kommer att användas.

1.1. Syfte och frågeställning

Det övergripande målet med CLIR är att möjliggöra en situation när en användare kan söka och återvinna dokument på olika språk genom att formulera och skriva in sin enda sökfråga på ett enda språk som han/hon behärskar bäst. I IR-sammanhang definierar man en sökfråga som en *query* d.v.s. en formell representation av ett informationsbehov i ett givet IR-systems språk. En query innehåller normalt ord, och kan även innehålla olika operatorer, t.ex. Booleska.

Ett av de centrala problem i CLIR-forskningen är hur man kan övervinna problem som queryöversättning medför och närma sig den återvinningseffektivitet som nås med den enspråkiga informationsåtervininningen. Syftet med denna magisteruppsats är att undersöka effekter av querystruktur i lexikonbaserad svensk-engelsk CLIR. Vi kommer att undersöka skillnaderna i återvinningseffektivitet mellan originalquery och översatt query och utvärdera hur querystruktur påverkar återvinningseffektivitet.

För att ta reda på detta har vi ställt följande forskningsfrågor:

1. Hur skiljer sig återvinningseffektiviteten mellan engelska originalqueries och lexikonbaserade engelska ostrukturerade queries, vilka har svenska som ursprung?
2. Hur skiljer sig återvinningseffektiviteten mellan engelska originalqueries och lexikonbaserade engelska strukturerade queries, vilka har svenska som ursprung?
3. Hur skiljer sig återvinningseffektiviteten mellan lexikonbaserade engelska ostrukturerade queries, som har svenska som ursprung, och lexikonbaserade, engelska strukturerade queries, som också har svenska som ursprung?
4. Hur ser resultaten ut i förhållande till tidigare forskning?

2. CLIR: centrala problem och metoder

2.1 CLIR som forskningsområde

Forskningen kring flerspråkig informationsåtervinning påbörjades så tidigt som på 1960-talet. Då utvecklades International Road Research system vilket använde kontrollerad vokabulär som kom från trespråkigt tesaurus med indexeringstermer på engelska, franska och tyska. År 1978 var en internationell standard om konstruktion av flerspråkiga tesaurer utarbetat, ISO Standard 5964. Flerspråkiga tesaurer kan dock inte lösa problemet fullständigt. Tre faktorer som höga kostnader, eftersläpning i aktualitet och svårighet med användning förknippade med tesaurer motiverade forskare att utveckla andra metoder.

Ett alternativ till användning av kontrollerad vokabulär är sökning i fritext med naturligt språk. Under 1990-talet rapporterades resultat av de första undersökningarna inom detta område i vilka man fastställde två huvudsakliga tillvägagångssätt: kunskapsbaserade och korpusbaserade (Oard & Diekema, 1998). Dessa tillvägagångssätt diskuteras i avsnittet om översättningsmetoder.

I tidigare forskning kring flerspråkig informationsåtervinning användes olika termer för fenomenet, bl.a., Multilingual Information Retrieval (Hull, 1996). Men från och med ASM SIGIR konferensen 1996 etablerades termen Cross-Language Information Retrieval. Hull (Ibid.) ger följande definitioner av ämnet CLIR:

1. IR på andra språk än engelska.
2. IR i samlingar med parallella dokument eller med dokument på flera språk där queryspråk bestämmer på vilket språk dokument ska återvinnas.
3. IR i enspråkig dokumentsamling i vilken en query kan köras på olika språk.
4. IR i en flerspråkig dokumentsamling där queries kan återvinna dokument på olika språk. Denna definition är en expansion av definition (3).
5. IR ifråga om dokument som innehåller delar på mer än ett språk.

Dessa definitioner beskriver CLIR med tanke på olika informationsbehov och berör olika matchningsstrategier som finns inom CLIR.

2.2 Matchningsstrategier

Man brukar skilja emellan fyra huvudsakliga matchningsstrategier: cognate matchning, queryöversättning, dokumentöversättning och interlingua-teknik.

Cognate matchning är en strategi som automatiserar processen i vilken man kan gissa betydelsen av en okänd term genom att se likheter i stavning och uttal. Cognate matchning används för termer som inte har någon översättningsekvivalens[1] och behålls oförändrade i queryn, t.ex. egennamn och domänspecifik terminologi. Denna strategi används ofta i kombination med andra strategier.

Queryöversättning är en mer utbredd strategi i vilken en query automatiskt översätts till språket som stöds av IR-systemet. Queryöversättningen är relativt effektiv men dess nackdel är att queries ofta är förhållandevis korta och består av några isolerade termer som saknar både grammatiskt och semantiskt sammanhang vilket ger upphov för flertydighet.

Dokumentöversättning är en strategi i vilken dokument automatisk översätts till alla språk som stöds av IR-systemet. Till skillnad från queryöversättning tillför den strategin mer kontext och minskar effekterna av översättningsflertydighet. Men översättning av en större dokumentsamling kan vara mycket resurs- och kostnadskrävande.

[1] Med *översättningsekvivalens* anses ett ord eller uttryck på målspråket som har använts i en viss text eller som man skulle kunna använda eller tänker använda som motsvarighet till ett ord eller uttryck på originalspråket (Ingo, 1991, s.84). Termen *översättningsekvivalens* förekommer i samband med lexikonbaserade metod (samt andra metoder) i CLIR, och betyder att man slår upp varje queryterm i en tvåspråkig ordbok eller ordlista och väljer en lämplig överensstämmelse mellan källspråk och målspråk.

Interlingua-teknik är en strategi med vilken både query och dokument kan översättas till en uniform språkoberoende representation. Ett vanligt exempel av denna strategi är tekniker som baserar sig på kontrollerad vokabulär. Eftersom varje term i kontrollerad vokabulär exakt motsvaras av bara ett begrepp kan termer från olika språk användas antingen för att indexera ett dokument eller för att formulera en query.

2.3 Översättningsmetoder

Man brukar skilja mellan kunskapsbaserade och korpusbaserade översättningsmetoder. Kunskapsbaserade metoder är sådana metoder som använder data som från början framställs och kodas manuellt, nämligen ontologier, tesaurer, tvåspråkiga maskinläsbara ordböcker samt maskinöversättningssystem (Oard & Diekema, 1998).

Ett vanligt sätt att översätta en query är att använda ett tvåspråkigt maskinläsbart lexikon (eng. MRD) för att slå upp alla termer och formulera en ny query med samtliga översättningsalternativ (Ibid.). Fördelen med MRD är att de är tillgängliga för många språkpar. Man kan också använda sig av ett tredje mellanliggande språk som ”pivot” språk om det saknas lexikon för något språkpar. Lexikon är relativt lätt att använda (jämfört med tesaurus), man behöver inte några större förkunskaper.

Nackdelen med lexikonbaserad översättning är att resultatet brukar påverkas av lexikonets omfång. De allmänna lexikonen innehåller de flesta ensamstående ord i ett språk. Men de brukar sakna sammansatta ord, facktermer och egennamn. Pirkola et al. (2001) definierar vidare problem med lexikonbaserad översättningsmetod som begränsningar av allmänna lexikon och oöversättbara söktermer, behandling av böjningsformer, frasigenkänning och -översättning och lexikalisk flertydighet i käll- och målspråket. Vissa problem kan man överbygga med hjälp av tesaurus.

Ett alternativt sätt för dokument- och queryöversättning är maskinöversättning (eng. MT). Maskinöversättningssystem strävar efter att överföra semantiskt förhållande mellan två språk på ord-, mening, snitt- och dokumentnivå. Idag finns flera kommersiella MT-system tillgängliga, bl.a. SYSTRAN, Globallink och EasyTrans. Men problemet med dessa system är att man inte kan påverka resultat av översättningen eftersom det är svårt att kontrollera vad som händer i MT-system. Systemet väljer inte fler översättningsalternativ utan bara ett översättningsalternativ och som eventuellt inte är rätt översättningsekvivalens. Felaktiga översättningar kan kraftigt påverka CLIRs återvinningseffektivitet. (Gey et al., 2003)

Korpusbaserade metoder innebär automatisk utvinning av översättningsinformation ur flerspråkiga testsamlingar (corpora) med hjälp av statistiska och matematiska metoder (Oard & Diekema, 1998). Man brukar ha i åtanke två typer av corpora: parallell och jämförbar.

Parallell corpora består av en dokumentsamling i vilken varje dokument är manuellt översatt till ett eller flera språk. Med dess hjälp skapar man två- eller flerspråkiga ordlistor som används för att bredda översättningsresurser för CLIR. Själva metoden består av att man drar ut en lista av nyckelord från dokumenten och sedan matchar orden på båda språken och får en ordlista som kan användas vid översättningen.

Jämförbar (eng. comparable) corpora är en tvåspråkig lingvistisk materialsamling som skapas via gruppering av liknande dokument med samma tema på två olika språk. T.ex. en utländsk utgåva av en nyhetstidning där artiklar om samma nyheter skrivs självständigt. Trots att denna teknik kan bidra till CLIR fick den inte så stort bruk. Orsaken är att data i ursprungliga texter inte är parallella p.g.a. att olika författare skriver med olika stilar och synvinklar samt kan texter med olika ämne placeras i samma domän. Dessutom kan texter, sådana som tidningsartiklar från olika tidsperioder, leda till variation i teman och göra tekniken svåranvändbar. (Gey, 2003).

2.4 Utvärdering av CLIR

Den del forskningen som är fokuserad på utvärdering av CLIR och är riktad till europeiska språk fick namnet Cross-Language Evaluation Forum (CLEF). CLEFs huvuduppgift är att utveckla en utvärderingsverksamhet för informationssökning på europeiska språk och konstruera en testmiljö samt utveckla forskningsmetoder som olika forskningsgrupper kan använda sig av för att få jämförbara resultat. Dessutom har CLEF som en viktig uppgift att skapa en forskningssamverkan inom CLIR-sektorn. Motsvarande uppgift har TREC Cross-Language track som riktar sig till arabiska språk och NTCIR Asian Language Evaluation Project som omfattar kinesiska, japanska och koreanska språk (Peters, 2000).

3. Viktiga lingvistiska fenomen inom IR och CLIR

De lingvistiska fenomen som studeras inom IR-området är: morfologisk variation, t.ex. avledning och böjning; sammansättning och frasigenkänning; lexikalisk flertydighet, t.ex. homonymi och polysemi (Hedlund, 2003, s.15). Dessa fenomen tillhör var för sig morfologisk, syntaktisk och semantisk sektion i lingvistiken.

Inom lingvistiken betraktas ett givet ord som en abstrakt enhet, ett lexem, som kan utföras av olika ordformer. Dessa ordformer kan vara böjningsformer eller avledningsformer. Morfologi är just den del av lingvistiken som studerar grammatisk struktur av ord och dess olika kategorier (genus, numerus, tempus, kasus, species: best./obest.) vilka representeras av ordformen samt ordbildning. Morfologisk analys delar ord till morfem (fri- och bundna morfem), d.v.s. de bildningsblock som ett ord består av.

Böjningen syftar på ett grammatiskt sätt att få fram olika ordformer av samma lexem och genom detta modifiera dess betydelse, t.ex., lexemet *nyhet* har följande böjningsformer: *nyheter, nyheterna.* Ändelserna ändrar betydelsen och ibland kan det spela roll om det är en eller flera nyheter eller om man talar om vissa speciella nyheter och inte vilka nyheter som helst (Bolander, 2001, s.101.). Det grammatiskt förhållande mellan dessa ordformer visar dock att *nyhet* är grundformen för samtliga.

Avledning är en form av ordbildning. Man lägger ett avledningsaffix (prefix eller suffix) till ett lexem som genom detta ger upphov till nya ord, t.ex. substantivet *generation* bildas av lexemet *genera* genom tillsättning av suffixet *-tion* medan plural av *generation* skaffas genom att lägga ändelsen *-er*.

Språkets morfologiska variation påverkar återvinningseffektivitet när sökterm skiljer sig formmässigt från motsvarande indexterm i databas då relevant dokument inte kan återvinnas. Detta blir mer komplicerat i CLIR, eftersom det inte bara kan röra om olika ordformer utan också om olika ordklasser eller till och med en fras som en översättningsekvivalens kan bestå av.

För att minska effekter av morfologisk variation använder man inom IR sådan metod som stamning (Pirkola et al., 2001). Stamning är en databehandlingsprocess under vilken böjnings- och ordbildningsaffixen tas bort och ord reduceras till sina stammar. Till exempel, *genera* är resultat av stamning för ordet *generation*. Det är inte nödvändigt att en sådan stam utgör ett riktigt ord. Principen som ligger bakom stamningen är att ord med samma ordstam är semantiskt relaterade (Moens, 2000, s.81) d.v.s. de har närliggande betydelse som representeras av en viss stam. Med andra ord kommer de ursprungligen från ett och samma lexem. Genom stamningen kan man undvika sökningar som missar relevanta dokument på grund av ordformens variation.

Effekterna av tillämpningen av en stamningsalgoritm har fått olika resultat. Baeza-Yates och Ribeiro-Neto (1999, s.168) påpekar att man inte är överens om stamningen bidrar till en väsentlig förbättring av återvinningseffektiviteten.

I ordböjningen kan oregelbundenhet förekomma vilket gör stamningen oeffektiv. T.ex., ordet *bok* som har stammen *bok* och i pluralis böjs som *böcker*, där stammen ändras till ”böck”. Därför kan stamningsalgoritm inte gruppera grundformen och böjningsformen tillsammans.

Normaliseringen är en process som omvandlar orden till dess lexikaliska grundform (Hedlund, 2003, s.20). T.ex. ord som *ran* i engelska normaliseras till *run*. Normalisering skiljer sig från stamning i och med att resultat av det förstnämnda utgör ett riktigt ord (grundform). Exempelvis för en term som *computations*, är resultat från stamning *comput*, medan produkten från normalisering är *computation*. (Ahlgren, 2004, s.18) Ordens lexikaliska grundform verkar vara av stor betydelse för lexikonbaserad CLIR (Hedlund, 2003, s.22).

Alla språk är inte uppbyggda på samma sätt, speciellt när det gäller sammansättning och fras. Ett och samma koncept kan utformas som sammansatta ord i vissa språk och som fras i vissa andra språk. Sammansättning är ett sätt att bilda nya ord genom att sätta ihop flera morfem som skulle kunna utgöra ett självständigt ord. Det är speciellt utmärkande för språk av den nordgermanska gruppen. Sammansatta ord i svenska motsvaras ofta av en fras i engelska. Sammansättningar och fras utgör ytterligare ett problem i IR och CLIR.

Pirkola et al. (2001) skiljer tre olika typer av sammansättningar: kompositionella och icke-kompositionella och semikompositionella. Betydelsen av ett kompositionella sammansatta ord kommer från dessa komponenter och är förutsägbar, t.ex. *barnbidrag= barn+bidrag*. Betydelsen av ett icke-kompositionella sammansatta ord framkommer inte tydligt av dessa komponenter och är inte förutsägbar, t.ex., *jordgubbar*. Splittningsmetoder (eng. decomposition methods) som man använder inom IR kan leda till olika konsekvenser som har stor betydelse för översättningsprocessen. Om man automatiskt splittar ett icke-kompositionellt sammansatt ord och översätter dess delar var för sig försvinner den rätta betydelsen av termen. Semikompositionella sammansättningar är sådana som delvis kan tolkas på basis av ordens komponenter men som också har metaforiskt eller en annan betydelse, t.ex. *krokodiltårar*.

En fras innebär en syntaktisk enhet som består av mer än ett separat ord. I en mening eller en sats måste dessa ord stå i en viss bestämd ordning i förhållande till varandra. Frasigenkänning vid automatisk översättning utgör ett betydlig större problem än sammansatta ord. Detta kan kraftigt påverka återvinningseffektivhet. För att lösa detta problem används olika frasigenkänningstekniker bl.a. statistisk metod, ordklasstagging och syntaktisk grundanalys (eng. shallow syntactic analisys) (Sheridan & Smeaton, 1992; Strzalkowski, 1995). I InQuery IR-system, som används för denna undersökning, markeras fras med avståndsoperatorn *#uwn* (unodered window n). InQuery samt funktionen av olika operatorer diskuteras närmare i metodkapitlet.

Det lingvistiska fenomenet som är av stor betydelse för CLIR är språkets lexikaliska flertydighet. Detta innebär att ett ord hänvisar till olika betydelser. Lexikalisk flertydighet orsakas av homonymi, polysemi och synonymi (Hedlund, 2003, s.19). Homonymi är ett sammanfall i uttal och stavning mellan olika ord, d.v.s. homonymer är olika lexem som har samma form. Betydelser hos homonymer är inte relaterade (Pirkola, 1999, s.24f). T.ex. *skär* (1) Hon *skär* brödet. (2) Hon strandade på ett *skär*. (3) Han har en *skär* slips.

Polysemi innebär ett förhållande där ett ord har flera, men oftast inte helt avvikande betydelser. Det kan vara ett gammalt ord som fått en ny eller utvidgad användning och betydelse (Bolander, 2001, s.34). Betydelserna är relaterade till varandra och kan komplettera varandra, bl.a. kan de ha metaforisk anknytning. T.ex. *stjärna*: (1) Hon såg en *stjärna* på kvällshimmeln. (2) Han var en *filmstjärna*.

Synonymi innebär att olika ord har nästan samma betydelse. Ord med exakt samma betydelse i alla kontexter, s.k. starka synonymer, förekommer väldigt sällan, t.ex. *både* och *bägge*. De flesta synonymer har samma betydelse bara i vissa kontexter. Det brukar alltid finnas en lätt skillnad mellan två ord som betraktas vara synonymer. Översättningsalternativen som finns förtecknade i lexikonet är mestadels synonymer. Synonymi kan påverka IR på två olika sätt. Eftersom olika författare har olika stilar och använder sig av olika termer för att beskriva samma ämne kan synonymer vara av stor hjälp. Man kan expandera query med synonymer för att öka effektivitet. Men vid översättningen kan synonymer dock bidra till ökningen av flertydighet.

Lexikalisk flertydighet har stora konsekvenser för CLIR. Den förekommer i både källspråk och målspråk. Den orsakar ökning av irrelevanta betydelser hos termer. En term kan ha en eller två betydelser i källspråket vilket vid översättningen kommer att uttryckas med flera översättningsalternativ med en utpräglad skillnad i betydelsen. Detta fenomen kallas för översättningsflertydighet. (Hedlund, 2003, s.19) Lexikalisk flertydighet i målspråket och översättningsflertydighet i källspråket resulteras i sämre återvinningseffektivitet.

3.1 IR-relaterade egenskaper hos svenska och engelska

Svenska och engelska tillhör den germanska språkgruppen och har en del liknande lingvistiska egenskaper. Men det finns ändå särskilda skillnader mellan svenskan och engelskan när det gäller det morfologiska systemet som är väsentligt för informationsåtervinningen. Svenskan har ett relativt komplext morfologiskt system som är rikt på böjnings- och avledningsmorfem. Det beror på att svenskan har ett utvecklat system som anger kategorier av numerus, genus och species hos substantiv och adjektiv, komparativa former hos adjektiv; samt tempus kategori hos regelbundna verb. Liksom engelskan har svenskan ett stort antal oregelbundna verb. Engelskan är relativt mindre morfologiskt komplext än svenskan när det gäller markering av sådana kategorier som numerus, genus och bestämd form.

Svenskan är dessutom mycket rik på sammansättningar. För att ange frekvens av sammansatta ord i svenskan har man gjort en analys av 100 000 ord från nyhetstidningar som visade att 9,8 % av alla ord var sammansatta ord (Hedlund, 2002). I engelskan motsvaras svenska sammansättningar ofta av fraser. Bägge språken har hög frekvens av homonymer och polysemer (Pirkola, 2001).

Vi ska dock inte ta upp alla lingvistiska fenomen som inte är direkt relaterade till vår undersökning utan kommer bara att diskutera sammansatta ord och dess översättning till engelska samt homonymi i bägge språken.

Svenska sammansättningar består vanligtvis av två eller flera stammorfem som binds ihop med eller utan hjälp av s.k. fogemorfem, t.ex., sammansättningen *gat**u**belysning* består av två stammorfem *gat-* och *belys-* och ett fogemorfem *u*.

Engelska sammansättningar kan skrivas på tre olika sätt: som en fras, t.ex. *air lane* (eng.) – *luftled* (sv.); som ett sammansatt ord, t.ex. *airline* (eng.) – *flyglinje* (sv.); och sammanskriven med bindestreck, t.ex. *air-conditioning* (eng.) – *luftkonditionering* (sv.)

Vid manuell översättning från svenska till engelska kan många sammansatta ord, som inte hittar exakt översättning i lexikon, effektivt splittas och översättas som separata ord i en fras. Splittningen kan göras med hjälp av lexikonbaserat morfologiskt program. Men översättning av icke-kompositionella sammansatta ord kan vålla problem eftersom ordets betydelse inte framgår av delarna efter splittningen. Men å andra sidan har vissa icke-kompositionella sammansättningar en exakt motsvarighet i engelskan, som t.ex. *jordgubbar* (strawberry). Pirkola (2001) påpekar dock att effekterna av sammansättningssplittning för IR inte är testade. Dessutom är det inte klart i vilken utsträckning svenska sammansättningar är genomskinliga.

Homonymi eller mer precis homografi (ord med samma grafiska form men med skilda semantiska betydelser) är ett väldigt vanligt fenomen i både svenska och engelska. Homografi kan påverka återvinningseffektiviteten likaså av originalquery som CLIR-query. Om man t.ex. vill hitta information om *en kanal* (vattenled) kommer man också återvinna information om en *TV-kanal*. Vid översättningen av en homograf får man ett stort antal irrelevanta termer i själva CLIR-queryn som kan resultera i lägre precision d.v.s. större antal irrelevanta dokument kan återvinnas.

4. Tidigare forskning

I detta kapitel ska tidigare experiment som står i samband med lexikonbaserad CLIR-forskning presenteras. Forskningen om lexikonbaserad CLIR har bedrivits under drygt ett decennium. För denna uppsats har undersökningarna valts för att de är knutna till centrala problem inom lexikonbaserad CLIR, och diskuterar lämpliga metoder som hanterar de berörda problemen som fortfarande är aktuella.

Det centrala problemet med lexikonbaserad CLIR är att queryöversättning leder till att återvinningseffektivitet sjunker med 40-60 % jämfört med enspråkig IR (Ballesteros & Croft, 1997). Forskare undersöker de faktorer som påverkar resultat och försöker att hitta metoder för att förbättra effektivitet. Bland de frågor som CLIR-forskare ställer är: hur kan lexikonens begränsningar övervinnas och hur kan oöversättbara termer hanteras? Hur kan man behandla morfologisk variation? Hur kan fras igenkännas och översättas och hur kan effekter av olika typer av frasöversättning hanteras? Hur kan lexikalisk och översättnings flertydighet bemästras?

Hull D. och Grefenstette G. (1996) genomförde experiment med franska queries och engelska dokument i TIPSTER testsamling. Syftet med undersökningen var att ta reda på vilka faktorer som påverkar återvinningseffektivitet av CLIR. För queryöversättningen användes ett tvåspråkigt transfererbart lexikon (eng. bilingual transfer dictionary). Utvärderingen av systemets effektivitet gjordes enligt följande strategi: Börja med queries, dokument och relevansbedömning på ett originalspråk, i detta fall engelska. Sedan översätts queries manuellt till ett annat språk, franska. Manuellt översatta queries översättas om med CLIR-system till engelska och sedan kan resultatet jämföras med originalqueries för att få fram bilden av prestation för CLIR-systemet.

För experimentet användes 50 TREC topics som manuellt översattes av professionella översättare. Vid queryöversättningen till CLIR-query gällde att varje ord i källspråket ersattes med alla möjliga definitioner i målspråket.

I sin undersökning har forskarna jämfört den original engelska queryn med tre versioner av CLIR-queries. Den första versionen var den automatiskt konstruerade queries med lexikon som beskrivs ovan. Den andra versionen använde en manuellt uppbyggd transfererbar ordlista som innehöll utvalda översättningar av querytermer. Den tredje version använde en manuellt uppbyggd transfererbar ordlista som innehöll frasöversättningar.

Resultatet av undersökningen visade att genomsnittlig precision av de ordbaserade CLIR-queries uppnådde ca 60 % av genomsnittlig precisionen av engelska originalqueries. De båda omarbetade ordbaserade och frasbaserade CLIR-queries uppnådde 68 % respektive 91 % av genomsnittlig precision av originalqueries vid DCV-nivåerna 5, 10, 15 och 20. Forskarna påpekar att frasigenkänning och frasöversättning är den viktigaste faktorn som påverkar återvinningseffektivitet av CLIR-queries även om översättningsflertydighet också bidrar till dålig prestation.

Hull D. (1997, 1998) undersökte problem med flertydighet som uppstår vid lexikonbaserad queryöversättning. Han genomförde ett spansk-engelskt CLIR-experiment med den *viktade Booleska* modellen i SMART IR-system. Den viktade booleska modellen tillåter att använda alla lexikonsynonymer utan risk att tillskriva för stor vikt till den översatta termen. Hull strukturerade queries och kombinerade alla översättningsekvivalenser av samma term med den Booleska operatorn *OR*. Han antog också att användning av *AND*-operatorn kan vara ett effektivt sätt att hantera flertydighetsproblem eftersom i målspråket korrekt översätta querytermer förekommer oftare tillsammans än i kombination av felaktiga översättningar.

I sin studie har Hull jämfört effektivitet av den viktade booleska modellen och vektormodellen. Resultatet visade att booleska strukturerade queries presterade bättre än vektorqueries speciellt när det gäller CLIR-problem. För bedömningen av återvinningseffektiviteten användes två mått: genomsnittlig precision för de 1000 första dokumenten och genomsnittlig precision för de fyra högsta precisionsvärden som ligger vid DCV-nivåer 5, 10, 15, 20. Genomsnittlig/högsta precision utgör 0,281/0,541 för booleska queries respektive 0,202/0,415 för vektorqueries. Men effekten av CLIR-queries strukturerade med booleska operatorer anses som relativt små: 0,304/0,568 för enspråkiga booleska queries och 0,281/0,541 för booleska CLIR-queries.

Pirkola A. (1998) har genomfört experiment med finsk-engelsk CLIR. Först översattes engelska originalqueries manuellt till finska. Sedan översattes queries från finska till engelska med hjälp av maskinläsbara lexikon och kördes i InQuery IR-system. Pirkola testade två querytyper i naturligt språk (NL): NL/Sentense, NL/Word-Phrase, som i sin tur var uppdelade i två undertyper: ostrukturerade och strukturerade. Dessutom har han testat tre queryövesättningsmetoder: *gd (general dictionary)* - med allmänt lexikon, *sd (special dictionary)? gd (general dictionary)* - med domänspecifikt lexikon och därefter med allmänt lexikon, *sd and gd* – med både domänspecifikt och allmänt lexikon.

Resultatet av hans experiment visar att det finns en väsentlig lucka mellan originalquery (BL) och ostrukturerade NL/S-queries. Vid 10 % recall utgör genomsnittlig precision av den ostrukturerade NL/S queryn 15,4 % och av originalqueryn för *gd*-queries - 37,9 % som är 22,5 % lägre än originalqueryn. Strukturering av NL/S-query visade en dramatisk förbättring av effektivitet. Vid 10 % recall visade den bästa CLIR-queryn, *sd and gd*, genomsnittlig precision på 35,9 % vilket bara är 2 procentenhet mindre än originalqueryn. För NL/WP-*gd* originalquery utgör genomsnittlig precision vid 10 % recall 31,8 % och för den ostrukturerade queryn 16,5 %. Genomsnittlig precision för den strukturerade NL/WP-*gd* queryn är 24,9 %. Pirkola påpekar att CLIR-queries baserade på maskinläsbar lexikonöversättning kan nå samma effektivitetsnivå som enspråkiga queries om man strukturerar query och använder både allmänt och speciellt lexikon vid queryöversättning. Genom detta kan problem med polysemi och lexikonbegränsningar övervinnas framgångsrikt.

Pirkola utvecklade vidare en querystruktureringsmetod och gjorde också experiment med andra språk. Experiment visade att struktureringen väsentligt kan förbättra återvinningseffektiviteten av långa queries samt förbättra effektiviteten av korta queries i lite minde

utsträckning. Dessutom påpekar Pirkola (1998) att användning av sökoperatorer *#syn* och *#uwn* i querystruktur är ett effektivt sätt att minska flertydighet.

Ballesteros L. och Croft, W. Br. (1996, 1997) har undersökt problem av frasöversättning för spansk-engelsk lexikonbaserad CLIR. De rapporterar att queries som översattes ord för ord visade en försämring av genomsnittligt precision på 55 % i jämförelse med originalqueries. 30 % av precisionssänkning anses vara p.g.a. översättningsflertydighet och 20 % p.g.a. dålig översättning av fraser. Forskarna antog att automatisk frasigenkänning och avgränsning skulle förbättra återvinningseffektiviteten. I deras experiment översattes spanska originalqueries till engelska basquery i vilken fras identifierades och markerades med ordklasstaggar.

Ballesteros och Croft (1997) har jämfört effektivitet av frasbaserade queries, där fras översattes som ett uttryck av flera termer, med effektivitet av ordbaserade queries, där fras översattes ord för ord, d.v.s. ordrelation till varandra inte markerades. Frasöversättningen (som uttryck) kan bli bra eller dålig. Undersökningen visade att en bra frasöversättning väsentligt kan förbättra resultatet. Genomsnittlig precision ökade med 150,3 %. Men en query med en dålig frasöversättning presterar värre än en ordbaserad query. Genomsnittlig precision minskar med 39,3 %. Resultatet antyder att översättningens kvalitet har större betydelse för frasbaserad översättning än för ordbaserad översättning. (Ibid, 1997, s.87)

Flertydighetsproblem kan hanteras genom omfattande queryexpansion. Tilläggstermer relaterade till det ursprungliga begreppet i queryn kan minska effekterna av felaktiga översättningar. Forskarna påpekar att pre-översättnings- och post-översättningsexpansion via automatisk feedbacktektnik förbättrar effektivitet av CLIR från 42 % till 68 % av enspråkigt återvinningseffektivitet (Ibid, s.90).

Hedlund T. (2003) undersökte i sin doktorsavhandling hur olika aspekter av svenska språket kan påverka flerspråkig informationsåtervinning. Hon beskriver det svenska språket som ett språk med en rik morfologi, med sammansättning som ett produktivt ordbildningssätt och med en relativt hög homonymfrekvens. I sin undersökning studerar hon bl.a. effekter av splittning av sammansatta ord. Splittningen har tillämpats när översättning av hela det sammansatta ordet inte finns med i lexikonet. Forskaren antar att översättning av komponenter av sammansatta ord som separata ord förmodligen leder till ökning av flertydigheten.

I ett av sina experiment testade Hedlund effekter av querystruktur för svensk-engelsk, finsk-engelsk och tysk-engelsk IR. Översättningarna gjordes automatisk med UTACLIR automatiska queryöversättnings- och querykonstruktionssystem. UTACLIR-systemet är fokuserat på språkspecifika egenskaper i både käll- och målspråket som morfologi och sammansättning. Det använder SweTWOL, FinTWOL, GerTWOL vilka är morfologiska program för normalisering av ordformer, stopordlista för borttagning av stoppord, normaliseringsverktyg för splittning av sammansättningar och borttagning av fogemorfem, och maskinläsbart tvåspråkigt lexikon för översättning. Systemet markerar oöversättbara ter-

mer, t.ex., egennamn, och placerar dem i oförändrad form i slutlig queryformulering. Ett antal probabilistiska och avståndsoperatorer används för querykonstruktion.

Testen visade att UTACLIR queryöversättningsmetod ger jämförbart resultat för alla tre källspråk och att återvinningsresultat var relativt stabilt. Prestation av ostrukturerade och strukturerade queries för alla språkpar pekar på att strukturerade queries presterar bättre än ostrukturerade queries. (Hedlund, 2003, s.52)

Adriani M. (2001) utförde forskning om engelsk-holländsk CLIR och studerade olika queryöversättningstekniker. Hon gjorde ett experiment i vilket hon kombinerade en lexikonbaserad teknik med en teknik som baserades på parallell corpora för att välja rätt semantisk betydelse som finns i lexikon för varje queryterm. Resultat visade att en ren lexikonbaserad teknik åstadkommer bättre queryöversättning än en teknik baserad på parallell corpora och den kombinerade tekniken.

De studier som presenterats i detta kapitel genomfördes med olika språk och använde olika struktureringsmetoder med vilka man försökte övervinna olika faktorer som påverkar resultatet. Trots skillnader i språk och metoder kan man jämföra de generella slutsatser som forskare har kommit fram till, genom att sammanställa resultat som nås med olika CLIR-metoder. Våra resultat kommer vi att jämföra framför allt med Pirkolas forskning (1998) som genomfördes i samma testmiljö och med samma struktureringsmetod som vi använt. Dessutom kommer vi att parallellisera våra iakttagelser med viktiga slutsatser av Hull (1997), Ballesteros och Croft (1997) och Hull och Grefenstette (1996).

5. Metod

I detta kapitel kommer vi att beskriva testmiljö, översättningsresurser samt experimentets tillvägagångssätt med konstruktion och översättning av queries, och testprocess. Dessutom kommer evalueringsmetod och relevanskriterier att behandlas.

5.1 Testmiljö

Testmiljön som används i undersökningen består av Query Performance Analyser, IR-systemet InQuery och en testkollektion som inkluderar 24 TREC topics och dokumentsamlingen TrecUTA.

5.1.1 QPA

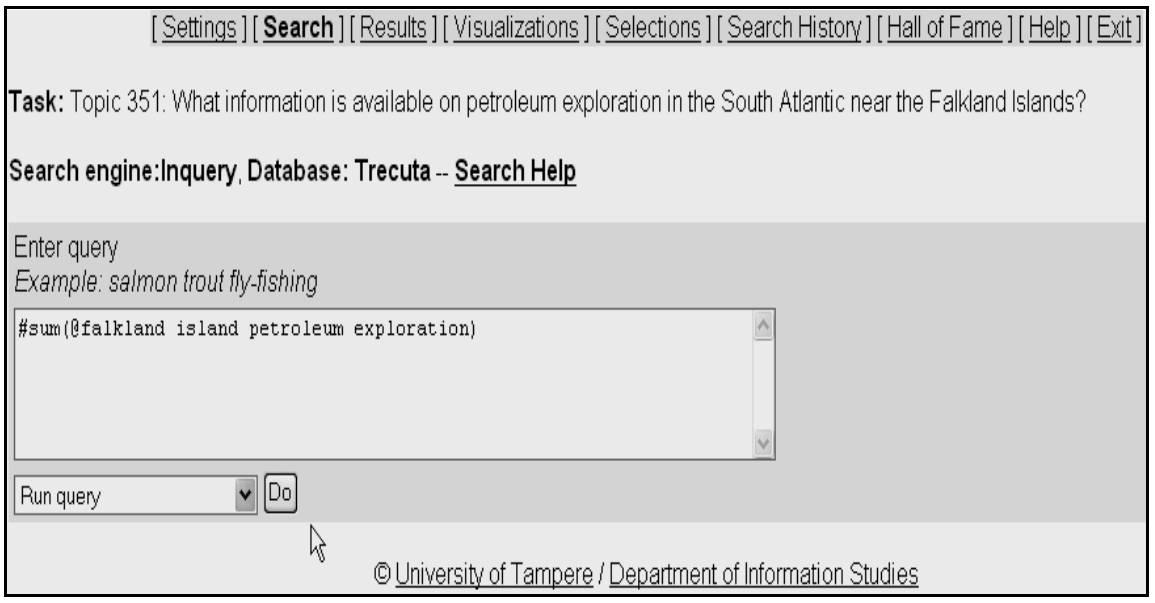
För att analysera och jämföra prestation av CLIR-queries kommer vi att använda Query Performance Analyser for IR-systems (QPA), version 5.1. QPA är ett testverktyg som är utvecklat av forskarna på Institutionen för Informationsvetenskap vid Tammerfors universitet för att möjliggöra en snabb analys av queryeffektivitet, samt jämförelse och visualisering av experiment inom IR inklusive CLIR. Detta verktyg innehåller en uppsättning av topics för sökning i text- eller bilddatabaser, relevansbedömningar som klargör vilka dokument som är relevanta för ett visst topic, en modul som stödjer queryformulering, användarsystem för att utföra queries i det utvalda IR-systemet samt en modul för inmatning av användarquery och visualisering av dess prestation (Sormunen, Halttunen & Keskustalo, 2002).

QPA tillåter att använda sådana yttre resurser som ett probabilistiskt IR-system InQuery, experimentdatabaser för IR (nu finns det flera databaser med engelska och finska dokument och en bilddatabas tillgängliga), tvåspråkiga elektroniska ordböcker som behövs för ordagrann översättning, samt morfologiska program för normalisering av ordformer.

QPA tillåter att analysera enskilda queries för att förstå variation i queryprestation och de orsakerna som ger upphov till denna variation genom att jämföra olika topics och olika queries. I CLIR-experiment kan variationen, t.ex., föranledas av sådana faktorer som syntaktiska fel i queryn, översättningsfel eller ”noisy” översättningstermer. Med QPA kan man snabbt göra korrigeringar av en query och se skillnaden i återvinningseffektivitet mellan den ursprungliga queryn och den korrigerade queryn.(Ibid.)

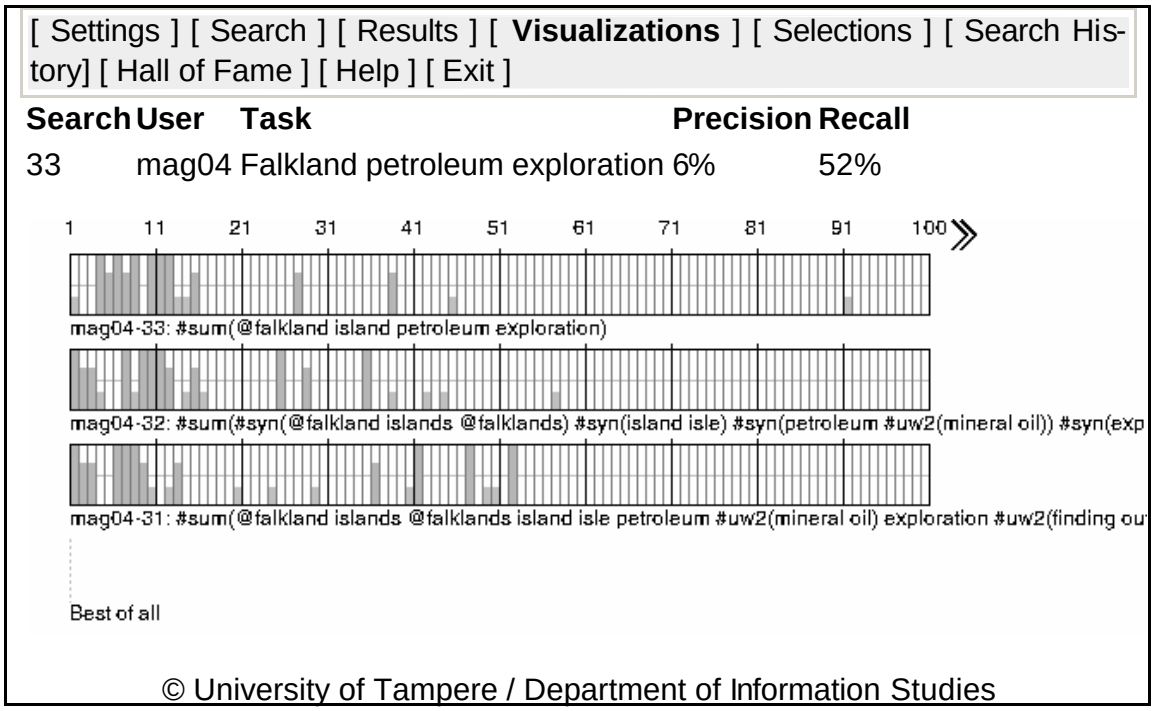
QPA användes för flera forskningsprojekt, bl.a. lexikonbaserade CLIR-experiment av A. Pirkola (1998) och andra forskare, och visade sig vara ett effektivt verktyg för att verifiera och förkasta forskningshypoteser.

Figur 1 illustrerar modulen för queryformulering och inmatning i QPA. Den består av topicbeskrivning, ett fält för queryinmatning samt en länk till hjälptext för queryspråk. Under queryinmatningsfältet finns ett funktionsfält där man kan välja bland tre funktioner: att köra query, att översätta query till ett urval språk eller att expandera query.



Figur 1. Fönster för queryinmatning i QPA.

Figur 2 illustrerar hur queryresultat kan visualiseras i QPA. Tre grafiska framställningar svarar mot tre olika querytyper: originalquery (mag04-33), en ostrukturerad query (mag04-31) och en strukturerad query (mag04-32). De nya versionerna av QPA (från 5.0) stödjer användning av större relevansuppgifter och tillåter att visualisera resultat med relevansskala som visar olika grad av relevans. Tre färger: grön, grå och vit används för att beskriva graden av relevans. Resultatet kan också framställas som Precision-Recall kurva. Dessutom kan en rankad lista över alla återvunna dokument visas och genom att klicka på dokumentets titel får man se dokumentet i fulltext.



Figur 2. Resultatvisualisering i QPA.

5.1.2 InQuery och querystruktur

Det IR-system som används i undersökningen är InQuery som ursprungligen var utvecklat vid Center for Intelligent Information Retrieval, University of Massachusetts. Det är ett probabilistiskt IR-system, d.v.s. ett system som baseras på sannolikhetslära, vilket implementeras med hjälp av det bayesianska nätverket. Genom samma grundtanke som ligger bakom den probabilistiska modellen försöker systemet estimera sannolikheten att ett dokument är relevant för en viss query vilket gör partiell matchning möjlig. Detta innebär att systemet tar hänsyn till osäkerhet, d.v.s. alla termer i en query behöver inte finnas i ett dokument för att dokumentet ska kunna återvinnas. InQuery mäter graden av likhet mellan dokumentet och queryn.

InQuery använder en termviktningsmetod d.v.s. varje term som ett dokument indexeras med har en vikt som avspeglar dess betydelse och utgör beliefvärde i dokumentet. Termviktningsmetoden diskuteras utförligt nedan.

InQuerys sökmotor stödjer två olika typer av query: ostrukturerad (eller svagstrukturerad) och strukturerad. En ostrukturerad query består av termsträng där inbördes förhållande mellan termerna inte markeras. I en strukturerad query begärs att en query matas i välstrukturerad format genom vilket manifesteras relationer mellan termerna.

För att formulera en query, erbjuder InQuery flertal olika sökoperatorer, vilka föregås av tecknet # och begränsas genom parenteser. I denna studie använder vi tre av dessa: *#sum*, *#syn* och *#uwn* som diskuteras nedan i samband med querystruktur.

Den senaste versionen av InQuery utnyttjar Kstem stamalgoritm för stamning och en rad morfologiska program för normalisering av indexeringstermer (Pirkola, 1998).

Querystrukturering innebär tillämpning av olika operatorer och parentes för att uttrycka relationer mellan querytermer (Kekäläinen & Järvelin, 1998, s.130.). Struktur hos en query kan vara svag eller stark.

En query med en svag struktur har relationer som anges av operator *#sum* som behandlar alla termer likadant. En query med en stark struktur har flera operatorer som differentierar förhållanden mellan querytermer, t.ex., *#sum* (*#syn*(a_1…a_n) *#uwn*(*bc*)). Operatorn *#syn* visar termernas semantiska förhållande till varandra och operator *#uwn* anger syntaktiskt förhållande mellan termerna.

Med hjälp av *#syn*-operator kan man strukturera en query via gruppering av synonymer genom att betrakta dem som instanser av samma begrepp och tillskriva större vikt till viktiga och korrekta termer.

#uwn (unordered window) är en avståndsoperator som tillämpas på fraser. Systemet återvinner bara dokumenten som innehåller argument av operatorn inom ramarna av ett avgränsat fönster (Pirkola et al. 2001), t.ex. *#uw2* anger att högst 1 ord får finnas mellan argumenten, som kan komma i godtycklig ordning. Om man vill söka dokumenten som handlar om eliminering av barnarbete på engelska kan man skapa en query: *#sum*(elimination *#uw2*(child labor)).

Querystruktur, som tillämpas i InQuery och anges genom användning av olika operatorer, hänvisar till beräkning av *beliefvärde* för varje enskild queryterm. Varje term som ett dokument indexeras med har en vikt som avspeglar dess betydelse och utgör beliefvärde i ett dokument. Beliefvärde *bv(k_i, d_j)* för ett term k_i med avseende på dokument d_i kan betraktas som vikten för k_i i dokumentet d_j och antas avspegla den betydelse som k_i har i d_i (Ahlgren & Eklund, 2004). För att en term ska få stor vikt i ett dokument ska följande förutsättning finnas: termen har en hög frekvens i dokumentet, dokumentet bör vara relativt kort jämfört med de andra dokumenten i samlingen och termen förekommer i ett litet antal dokument i samlingen (Ibid.).

Beliefvärde beräknas enligt följande ekvation:

$$bv(k_i, d_j) = 0.4 + 0.6 * \left(\frac{tf_{ij}}{tf_{ij} + 0.5 + 1.5 * (dl_j / adl)} \right) * \frac{\log((N + 0.5)/df_i)}{\log(N + 1.0)} \qquad \textbf{(1)}$$

Där

tf_{ij} = frekvensen (antalet förekomster) av k_i, i d_j
dl_j = längden på d_j (antalet ordförekomster i d_j)
adl = den genomsnittliga dokumentlängden i D
N = antalet dokument i D
df_i = antalet dokument i D i vilka k_i förekommer.

De termer som inte förekommer i dokumentet d_j tilldelas värdet 0,4 och inte 0. Följaktligen blir det lägsta värdet = 0,4 och ekvationen (1) ger att $0.4 = bv(k_i,d_j)<1$. Denna termviktningsmetod, som används i InQuery IR-system, är en modifiering av *tf-idf*-metoden[2].

Förutom termviktningen, beräknar InQuery även ett beliefvärde för hela queryn i relation till dokumentet, vilket kan betraktas som InQuerys *likhetsvärde*. *Likhetsvärde* berör dokument med avseende på hela queryn q i relation till dokumentet d_j (Ibid.). Detta är utgångspunkten för rankning av databasens dokument.

Beliefvärdet för en query med en svag struktur som representeras av #*sum*-query med avseende på ett dokument d_j utgörs av medelvärdet över beliefvärde för varje term som består av antigen ett ord eller en fras (en fras betecknas som #*uwn*-uttryck). Beliefvärdet för en #*sum*-query definieras så här:

$$bv(\#sum(E_1, E_2, \ldots E_n), d_j) = (p_1+p_2+\ldots+p_n)/n$$

där p_i är beliefvärdet för termer eller operatoruttryck E_i med avseende på dokument d_j.

Beliefvärdet för en query med en stark struktur som representeras av en #*syn*-query med avseende på ett dokument d_j och med uteslutande termer som operander (det spelar ingen roll hur många av termerna inom #*syn*-operatorn som är närvarande i dokumentet d_j utan beror på termernas frekvens i dokumentet d_j) och definieras enligt ekvation (2):

$$bv(\#syn(E_1, E_2, \ldots E_n), d_j) = 0.4 + 0.6 * \left(\frac{\sum_{k_i \in S} tf_{ij}}{\sum_{k_i \in S} tf_{ij} + 0.5 + 1.5 * (dl_i / adl)} \right) * \frac{\log((N + 0.5) / df_S}{\log(N + 1.0)} \qquad (2)$$

där S är mängden av termer inom #*syn*-operatorn, tf_{ij} frekvensen för termen k_i i d_j och df_S antalet dokument i D som innehåller minst en av termerna i S. I produktens andra komponent summeras frekvenserna i d_j för samtliga termer inom #*syn*-operatorn. (Ahlgren & Eklund, 2004).

Denna termviktningsmetod har stor betydelse för CLIR. I lexikonbaserad CLIR begränsas man inte till ett översättningsalternativ utan har samtliga översättningar som finns med i lexikonet och plockas fram vid automatisk översättning. Problemet uppstår när man väljer bland olika termalternativ och viktar dem (Hedlund, 2003, s.30). Om man exempelvis har en originalquery som består av tre termer #sum(A, B, C) där alla har samma vikt och översätter den första termen med två motsvarigheter $A?$ a_1, a_2, den andra med

[2] Se termviktningsmetod för IR-vektormodellen i Baeza-Yates & Ribeiro-Neto (1999, s. 29-30)

fyra *B?* b_1,b_2,b_3,b_4 och den tredje med en *C?* *c* kommer en ostrukturerad CLIR-query att se ut enligt följande:

$$\#sum(a_1, a_2, b_1, b_2, b_3, b_4, c)$$

På så sätt lägger man slumpmässigt större vikt till termer som har fler översättningsalternativ.

Pirkola (1998) föreslog att gruppera översättningsmotsvarigheter av samma begrepp med #*syn*-operatorn. Denna operator betraktar sina operander som instanser av samma begrepp. I enspråkig IR rör det sig om morfologiska varianter av samma ord eller olika termer för samma begrepp, till exempel synonymer. I CLIR gäller det översättningsvarianter, d.v.s. lexikonets synonymer till ett uppslagsord:

$$\#sum(\#syn(a_1,a_2)\ \#syn(b_1 \ldots b_4)\ c)$$

där a_1, a_2, $b_1 \ldots b_4$, *c* står respektive för översättningsmotsvarigheter av termerna *A, B, C*. Med #*syn*-operatorn grupperas lexikonets synonymer, d.v.s. översättningsmotsvarigheter som lexikonet grupperar tillsammans, och beliefvärdet av synonymgruppen beräknas enligt ekvation (2).

Dessutom kan man inte vid automatisk översättning skilja mellan synonymer och homografer. Detta försöker man också överbrygga med strukturerad översättning genom att separera olika grupper av översättningsvarianter vilka anges av lexikonet.

5.1.3 Testkollektion

I vår undersökning kommer vi att använda TrecUta databas som är inbyggd i QPA. Denna textkollektion, som alla andra i TREC, består av tre delar: dokumentsamling, topics och en mängd relevanta svar. TrecUta innehåller ca 550 000 dokument som består av engelskspråkiga artiklar från amerikanska tidningar: the Financial Times Limited 1991-1994, the Congressional Record of the 103rd Congress (1993), Federal Register (1949), the Foreign Broadcast Information Service (1996) och the Los Angeles Times (1989-1990). (InQuery/TREC-Uta search Guide)

Databasen är grundformindexerad med hjälp av EngTWOL, ett program för morfologisk analys av engelska ord. Till databasen tillhör 41 topics som finns numrerade (25-65) i QPA och ett antal relevanta dokument för varje topic. För vår undersökning har vi valt att testa 24 topics för vilka vi kan tillämpa *syn*-baserad querystrukturering.

5.2 Översättningsresurser

I denna studie, använder vi Nordstedts tredje upplaga av *Den stora engelska online ordboken*. Ordboken består av Stora *Svensk-engelska ordboken* (SSEO1988) och Stora *Engelsk-svenska ordboken* (SESO 1980).

Den stora engelsk-svenska ordboken är Sveriges första datoriserade tvåspråkiga ordbok och är också en av de första i världen. Ordboken omfattar ca 120 000 ord och fraser. Urvalsprincipen täcker det moderna brittiska och amerikanska engelskans allmänna ordförrådet. Det tar också hänsyn till formellt och äldre språkbruk, samt viktiga facktermer inom ämnen som affärsliv, medicin, teknik, sport mm. Den svensk-engelska ordboken bygger på Engelsk-svenska ordbok (1980) och har liknade principer och omfång.

Vi har använt oss av *WordFinder* software som innehåller maskinläsbara tvåspråkiga Nordstedts engelsk-svenska och svensk-engelska lexikon.

Ordboksartiklarna indelas på flera sätt. Ord med samma stavning men med olika ursprung och betydelse (homografer) är uppställda som olika uppslagsord med fet arabisk siffra framför:

> **1 förband** s. **1** med. bandage; kompress o.d. dressing; **2** mil. unit; flyg. formation
> **2 förband** s. mus. warm-up band

Med liten stil anges ämnesområde eller andra förklaringar samt delbetydelser hos översättningar. Arabiska siffror som står efter uppslagsord delar upp i betydelser:

> **(blåsa** s **1** i huden, i metall, glas, målning blister, i glas äv. bleb … **2** anat., isht el. luftbehållare bladder, vetensk.äv. vesica **3** bubbla bubble **4** vard., festklänning party dress)

I vår undersökning kommer vi att betrakta som synonymer alla översättningar som står uppräknade efter ett uppslagsord utan siffror samt de som är uppräknade efter varje enskild siffra som är uppsatt efter uppslagsord. Synonymer kommer att grupperas tillsammans. Översättningarna som anges av lexikonet under olika siffror efter ett uppslagsord har större skillnader i betydelse än synonymer och därför kommer att placeras i skilda grupper. Översättningarna som anges efter olika uppslagsord (homografer) kommer också att placeras i separata grupper. Exempelvis kommer översättningsalternativen i den strukturerade queryn att markeras med *#syn*-operator och stå i följande ordning:

> *#sum*(*#syn*(bandage dressing) *#syn*(unit formation) *#uw2*(warm-up band))

5.3 Tillvägagångssätt

5.3.1 Val av topics

Av de 41 topics som tillhör databasen TrecUta har vi valt 24 topics. Eftersom vi använder ett allmänt lexikon är ämnet, som behandlas i topicsen, av ringa betydelse för vår undersökning. Urvalet baseras på förutsättningen att en query som vi får efter översättningen innehåller fler termer än den ursprungliga queryn för att strukture ningen ska kunna tillämpas. För topics i vilka varje queryterm motsvarades av bara en översättningsterm, kan struktureringen inte tillämpas, och därför de valdes bort.

5.3.2 Konstruktion och översättning av queries

I vår studie kommer vi att formulera för varje av de utvalda 24 TREC topics tre queries: en originalquery och två CLIR-queries: en ostrukturerade och en strukturerade.

En *originalquery* är en query som består av termer i grundform hämtade från ett topic. T.ex., för topic 29: What are the benefits, if any, of drug legalization? kommer en originalquery se ut så här: *#sum*(drug legalize benefit)

En *ostrukturerad* query är en söksträng, där alla querytermer sätts ihop med operatorn *#sum* och fraser, som representerar en term, kan markeras med avståndsoperatorn *#uwn*. Varje term får samma vikt i queryn. Inbördes förhållande mellan termerna markeras inte, d.v.s. lexikonets synonymer grupperas inte med hjälp av *#syn*-operatorn. T.ex., den ostrukturerade CLIR-queryn för det ovannämnda topicet 29: *#sum*(drug legalize authenticate adva ntage benefit profit advantage vantage van)

En *strukturerad* query är en söksträng i ett välstrukturerad format som i sin tur manifesterar relationer mellan söktermerna. Man grupperar alla lexikonets synonymer och markerar dem med *#syn*-operatorn. Varje grupp av termer får samma vikt. En strukturerad CLIR-query för det ovannämnda topicet 29 kan se ut så här: *#sum*(drug *#syn*(legalize authenticate) *#syn*(advantage benefit profit advantage vantage van))

Konstruktions- och översättningsprocess kommer att se ut på följande sätt:
Först, konstruerar vi en engelsk query som kommer att tjäna vidare som en originalquery. Den kommer att bestå av de viktigaste orden och fraser från topictexten, vilka motsvarar Pirkolas (1998) NL/WP-queries. Termer ska bindas ihop med hjälp av operator *#sum*. Fraser hålls ihop med avståndsoperatorn *#uwn*.

Vidare kommer vi att manuellt översätta ord/fraser i originalqueries till svenska med hjälp av online Nordstedts engelsk-svenska lexikon och andra lexikon och kontrollera översättningarna i Svenska Akademins Ordlista. Den rätta översättnings motsvarigheten av det engelska ordet ska väljas med hjälp av den kontext, som ges av topictexten. T.ex, kommer ordet *"effect"* i följande topictext: "What effects have been attributed to El Nino?**"** att översättas med *"effekt"* eller *"verkan"* men inte med *"innehåll"* som är en av

möjliga översättningar av ordet *"effect"*. De resulterande svenska orden tranformeras till grundform.

De engelska fraser som finns med i Nordstedts engelsk-svenska lexikon kommer att markeras med avståndsoperator #uw2 och ska översättas med motsvarande svenska uttryck eller sammansatta ord, t.ex. #uw2(*south american) ? sydamerikan.* Om en engelsk fras inte finns med i lexikonet kommer den att översättas ordagrant utan att skrivas ihop, t.ex. *blood alcohol level* ? *blod alkohol halt*.

Sedan ska vi fortsätta med att simulera automatiskt översättning av den svenska queryn. Därför är det lämpligt att ta med samtliga engelska ord som lexikonet ger. Med hjälp av online Nordstedts svensk-engelska lexikon kommer vi att översätta den svenska queryn till engelska och skapa med alla översättningsvarianter en ostrukturerad CLIR-query. Här tillämpas *#uw2* på engelska fraser som ges av lexikonet.

I enlighet med Pirkolas studie (1998) grupperar vi lexikonets synonymer till ett givet svenskt uppslagsord inom *#syn*-operatorn och skapar den tredje querytypen - en strukturerad query.

Om det saknas en engelsk motsvarighet av ett svenskt ord kommer det svenska ordet att finnas i den nya engelska queryn. Tecknet "@" kommer att sättas framför dessa ord. I den engelskspråkiga databasen kan en del termer på andra språk finnas i indexet med tecknet "@" framför. Detta gäller framförallt egennamn och ämnesspecifika termer (Ahlgren, 2005), t.ex. @osteoporosis. Problemet uppstår delvis också på grund av att vissa nya termer inte är betecknat i det allmänna tvåspråkiga lexikonet. T.ex. är ordet "kodning" en översättning till ordet "encryption" i topic 34: "Identify documents that discuss the concerns of the United States regarding the export of encryption equipment". I online Nordstedts svensk-engelska lexikon saknas idag översättning till ordet "kodning". Därför kommer vi att placera det svenska ordet i CLIR-query i samma form och testa om det eventuellt finns i indexet föregånget av tecknet "@".

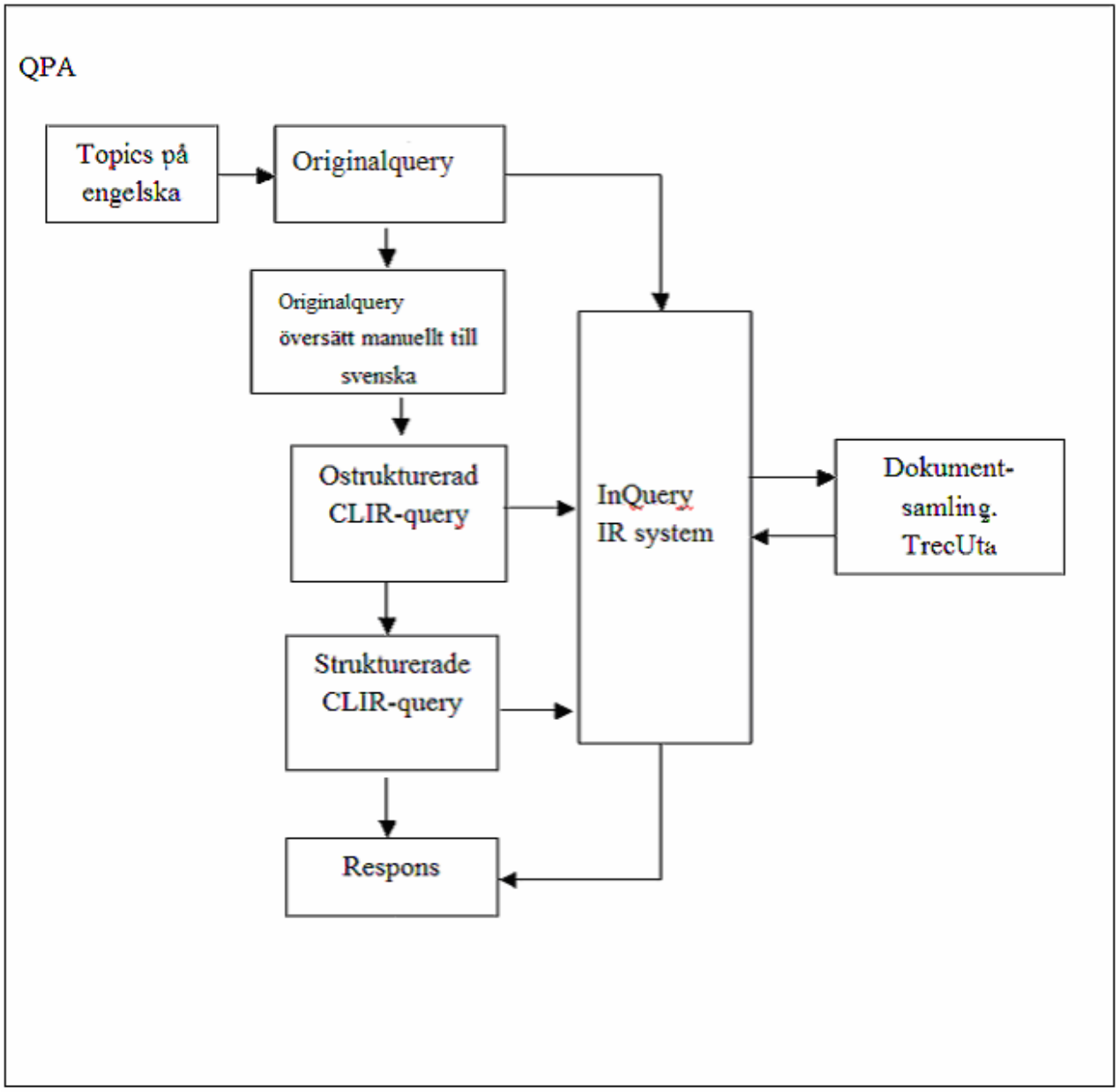
Eftersom TrecUta databas är grundformindexerad kommer vi att testa varje engelsk queryterm *t* (grundform) om ordet har känts igen eller inte av EngTWOL. Om inga dokument återvinns när vi kör *t* kommer vi att använda "@" före den termen också. I fall om minst ett dokument återvinns, ska vi använda term *t* i queryn. T.ex., i topic 26: "Identify systematic explorations and scientific investigations of Antarctica, current or planned" återvinns inga dokument när vi söker med *antarctica*, och därför ska *@antarctica* användas i stället.

5.3.3 Testprocess

Testdata i vår undersökning kommer att skaffas av tre querytyper. För varje querytyp kommer 24 topics att testas. Sammanlagt kommer vi att göra 72 querykörningar och få 72 rankade resultatlistor av dokument. DCV=100 kommer att användas för resultatlistor.

Först kommer vi att köra den engelska originalqueryn mot systemets databas för att få återvinningseffektivitet av enspråkig sökning. Sedan kommer vi att köra en ostrukturerad och en strukturerad CLIR-queries och jämföra resultat med originalqueryn. Data kommer att insamlas i en Excel-fil och bearbetas vidare.

Undersökningsprocess illustreras med Figur 3.



Figur 3. Undersökningsprocess

5.4 Evalueringsmetod

Precision och *recall* är två mått som traditionellt används inom IR för utvärdering av återvinningseffektivitet. *Precision, P*, används för att ange systemets förmåga att hämta relevanta dokument och definieras som antalet återvunna relevanta dokument, *a*, dividerat med samtliga återvunna dokument, relevanta, *a* och irrelevanta, *b:*

$$P = \frac{a}{a+b}.$$

Recall är ett mått som mäter fullständighet hos mängden av relevanta dokument som systemet returnerar med avseende på samtliga relevanta dokument som finns i databas för en respektive topic. *Recall, R*, definieras som antalet återvunna relevanta dokument dividerat med antalet alla relevanta dokument som finns i databasen för en respektive topic, återvunna, *a*, och icke-återvunna, *c*:

$$R = \frac{a}{a+c}.$$

I vår studie kommer vi att använda *genomsnittlig precision* (medelvärdet av precisionsvärden) vid olika DCV-nivåer. Detta mått visar inte bara hur många relevanta dokument som återvunnits utan också tar hänsyn till hur högt de relevanta dokumenten placerar sig i en träfflista. Det erhållna precisionsmåttet visar inte hur värdet påverkas av de relevanta dokumentens position i träfflistan.

Vi antar, exempelvis, att en query återvinner tre relevanta dokument inom ett DCV på 5 och dessa befinner sig vid position 1, 2, 3 och en annan query återvinner också tre relevanta dokument men vid position 3, 4, 5. Den erhållna precisionen för båda queries blir 3/5= **0.6** medan den genomsnittlig precision för den första queryn blir (1/1+2/2+3/3+3/4+3/5)/5= **0,87** och för den andra queryn blir (0/1+0/2+1/3+2/4+3/5)/5= **0,286667**.

Med hjälp av genomsnittlig precision kan man skapa en precisionskurva som visar genomsnittligt precision vid varje DCV-nivå för varje typ av query.

För att utvärdera systemets återvinningseffektivitet kommer vi att jämföra effektiviteten för ostrukturerade CLIR-queries och strukturerade CLIR-queries samt jämföra resultat för CLIR-queries med resultatet för enspråkig originalquery. För detta kommer *genomsnittlig precision* att beräknas över alla topics för DCV-nivåer 5, 10, 20, 30, 50, 60, 70, 80, 90, 100 och för varje querytyp.

I syfte att jämföra IR-metoder (querytyper) ”topic-by-topic” kommer vi att beräkna genomsnittlig precision över alla använda DCV-nivåer för ett givet topic och en given querytyp. Resultatet kommer att presenteras i tabeller och diagram.

5.5 Relevanskriterier

Det finns olika kriterier som man kan använda för att göra relevansbedömningar, bl.a. systemrelevant, ämnesrelevant, kognitivrelevant, osv. För vår studie kommer vi att använda ämnesrelevans. Det innebär att ett dokument anses relevant om det *handlar om* det önskade ämnet som uttrycks i queryn (Ahlgren, 2004, s.39). Ämnesrelevans grundar sig just på dokumentets ”aboutness” d.v.s. en query som *mercy killing* återvinner dokument som handlar om eutanasi som ämne. Det skiljer ämnesrelevans från systemrelevans som baseras på termmatchning, d.v.s. termer som förekommer både i queryn och i dokumentet, t.ex., samma query kan återvinna dokument som innehåller lösa termer *killing* eller *mercy* men inte handla om *eutanasi.* Därför betraktas de dokument som **inte** handlar om ämnet som icke-relevanta.

De flesta TREC och CLEF testkollektioner som används inom IR forskning tillämpar binär relevansbedömning med relativt liberala relevanskriterier. Binärrelevansbedömning har följande fördelar (Sormunen, 2002):

- den gör beräkning av prestationsresultat relativt enkel;
- är kostnadseffektiv;
- garanterar mätningsstabilitet genom att maximera antal relevanta dokument per topic.

Men den binära relevansskalan har fått mycket kritik på senare tiden. IR-system som effektivt återvinner högrelevanta dokument kan drabbas av binär relevansbedömning. Dessutom saknades tydliga definitioner av relevanskriterier (Ibid.).

I sin artikel rapporterar Sormunen (2002) resultaten av ett projekt vars syfte var att analysera och karakterisera poolen med relevanta dokument samt definiera relevanskriterier. Relevanskriterier måste visa skillnad mellan dokument som innehåller mycket information om ämnet (högrelevanta och relevanta) och dokument som innehåller lite ämnesinformation (marginellt relevanta). Därför tillämpades en relevansskala med fyra möjliga poänger: 0 – icke-relevant, 1 – marginellt relevant, 2 – relevant, 3 – högrelevant. Ett dokument betraktas som:

- *icke-relevant (0)* om det innehåller ingen relevant information för gällande topic;
- *marginellt relevant (1)* om det bara hänvisar till ämnet och inte innehåller mer information än vad som framgår från topicbeskrivning, vanligen en mening eller några få uppgifter som är relevanta;
- *ganska relevant (2)* om det innehåller mer information än gällande topic men inte uttömmande, vanligen ett relevant stycke eller 2-3 meningar;
- *högrelevant (3)* om det innehåller relevant och uttömmande information för gällande topic, vanligen flera relevanta stycken eller mer än fyra relevanta meningar.

Fördelen med en fyrgradig relevansskala är att det kan göras åtskillnad i viktning mellan dokument som är relevanta och nyttiga (rel.=2) och dokument som är marginellrelevanta (rel=1) men potentiellt icke-nyttiga (Ibid.).

Dokument som ingår i vår studie har relevansbedömts enligt den fyrgradiga skalan som finns inbyggd i QPA. Men eftersom antalet av högrelevanta och relevanta dokument tenderar att blir relativt litet kommer vi dock i vår studie att använda oss av ett binärt system och betrakta dokumenten som får 0 som icke-relevanta och dokumenten som får 1, 2, 3 som relevanta för att få stabilt resultat.

6. Resultatredovisning och analys

I detta kapitel kommer vi att presentera undersökningens resultat, analysera det erhållna resultatet samt besvara forskningsfrågorna. I det följande kapitlet jämför vi vårt resultat med tidigare forskning.

6.1 Resultatredovisning

6.1.1 Precision vid olika DCV-nivåer

För 24 testade topics och för varje querytyp har vi beräknat genomsnittlig precision vid elva DCV-nivåer 5, 10, 20, 30, 40, 50, 60,70, 80, 90, 100. Därefter har medelvärde för samtliga topics för respektive querytyp och DCV-nivå redovisas i Tabell 1 och Diagram 1. För varje enskilt topic har vi beräknat medelvärde över alla använda DCV-nivåer för respektive querytyp (Tabell 2, Diagram 2). I tabellerna och diagrammen använder vi oss av följande förkortningar: *original.* för originalquery, *ostrukt.* för ostrukturerad query, *strukt*. för strukturerad query, *MV* för det sammanfattande medelvärdet över värden i tabeller.

Tabell 1 och Diagram 1 visar att originalquery presterar bäst vid DCV-nivåer 5, 10, 20,30, 90. Strukturerad query har lite lägre precision i början men från och med DCV=40 visar strukturerad CLIR-query det bästa resultatet med ett undantag på DCV=90. Ostrukturerade query visar lägst precision vid alla använda DCV-nivåer.

DCV-nivå	*Original.*	*Ostrukt.*	*Strukt.*
5	0,442	0,333	0,417
10	0,388	0,263	0,350
20	0,308	0,209	0,306
30	0,269	0,178	0,257
40	0,235	0,164	0,236
50	0,213	0,148	0,220
60	0,182	0,139	0,199
70	0,179	0,127	0,183
80	0,167	0,117	0,170
90	0,187	0,108	0,160
100	0,148	0,101	0,152

Tabell 1. Medelvärde över alla topics för tre querytyper vid olika DCV-nivåer

Diagram 1 visar att den genomsnittliga precisionen av originalqueryn höjs en del vid DCV=90 medan den genomsnittliga precisionen av de båda CLIR-queries sjunker stabilt.

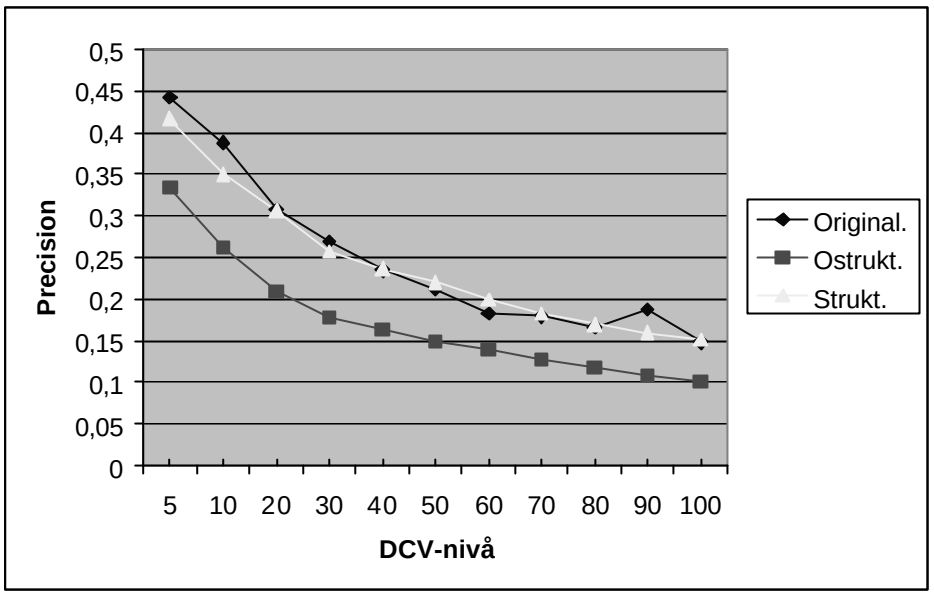


Diagram 1. Medelvärde över alla topics för tre querytyper vid olika DCV-nivåer

6.1.2 Precision för enskilda topics

Tabell 2 visar medelvärdet över de använda DCV-nivåerna för varje enskilt topic och för de tre querytyperna. Medelvärden över samtliga topics visar att det finns en skillnad på 8 procentenheter mellan originalqueryn och den ostrukturerade queryn. Mellan de ostrukturerade och strukturerade queries föreligger en skillnad på 7 procentenheter. Originalqueryn fick det högsta värdet, 25 %, medan den ostrukturerade queryn fick det lägsta värdet, 17 %. Skillnaden mellan originalqueryn och den strukturerade queryn är marginell, och utgör endast 1 procentenhet.

Topic	*Original.*	*Ostrukt.*	*Strukt.*
25	0,360	0,321	0,354
26	0,363	0,031	0,409
28	0,118	0,258	0,269
29	0,802	0,716	0,813
31	0,546	0,246	0,535
34	0,190	0,000	0,000
36	0,070	0,027	0,023
39	0,181	0,026	0,137
40	0,211	0,070	0,182
42	0,000	0,045	0,021
43	0,375	0,353	0,381
45	0,519	0,414	0,472
46	0,217	0,133	0,175
47	0,084	0,085	0,171
48	0,179	0,177	0,192
49	0,034	0,018	0,082

50	0,318	0,334	0,355
56	0,354	0,140	0,321
57	0,006	0,019	0,010
58	0,228	0,053	0,195
61	0,047	0,045	0,060
62	0,269	0,332	0,332
64	0,275	0,192	0,175
65	0,183	0,081	0,117
MV	0,247	0,172	0,241
MV,%	25%	17%	24%

Tabell 2. Genomsnittlig precision över elva DCV-nivåer för varje enskilt topic och för de tre querytyperna samt medelvärdet över samtliga topics.

Diagram 2 visualiserar medelvärden över de elva DCV-nivåerna för varje enskilt topic. Diagrammet visar att originalqueryn presterar bäst för hälften (tolv) av topics (25, 31, 34, 36, 39, 40, 45, 46, 56, 58, 64, 65). Den ostrukturerade CLIR-queryn har högst medelvärde för två topics (42, 57). Den strukturerade CLIR-queryn har fått det bästa resultatet för nio topics (26, 28, 29, 43, 47, 48, 49, 50, 61). För topic 62 har den ostrukturerade och den strukturerade queryn samma värde, som är större än värdet för originalqueryn.

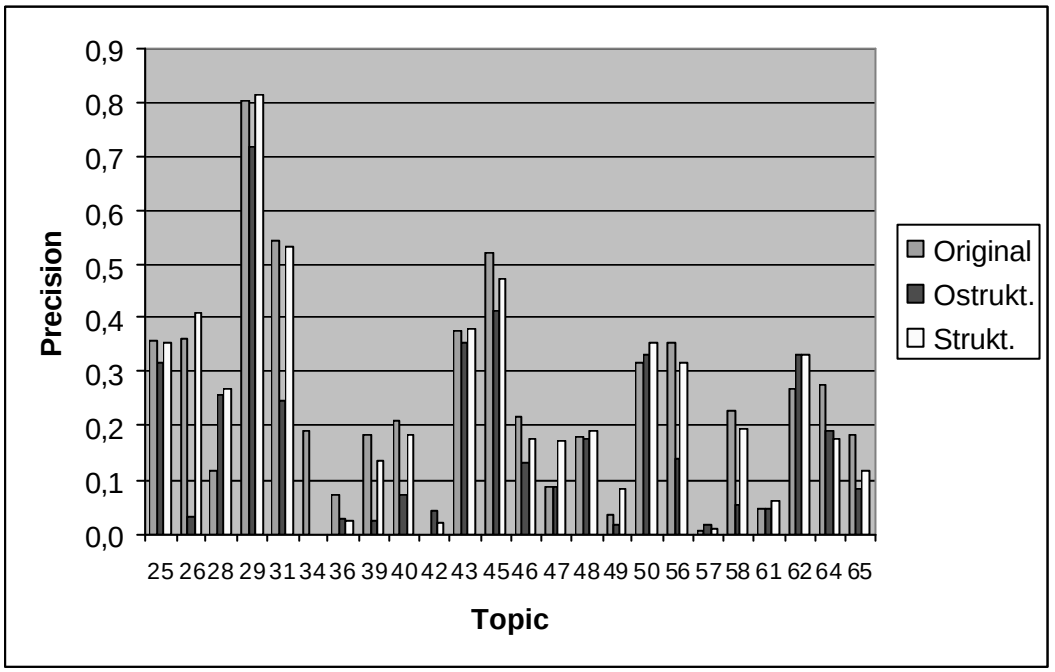


Diagram 2. Genomsnittlig precision över alla DCV-nivåer för varje enskilt topic och för de tre querytyperna.

6.1.3 Precision, topic för topic

Prestation över samtliga topics vid olika DCV-nivåer visar bara en del av bilden, därför redogör vi nedan prestation över enskilda topics. Av de 24 topics som undersöktes visar medelvärdena över de använda elva DCV-nivåer (Tabell 2) att originalqueryn fick det högsta värdet i 12 topics som utgör 50 % av alla använda topics. Den ostrukturerade queryn fick det högsta värdet bara i 2 topics, 8,3 %. Den strukturerade queryn fick det högsta värdet i 9 topics, 37,5 %. 18 topics, 75 %, visade förbättrat resultat i den strukturerade queryn jämfört med den ostrukturerade queryn.

För att visa i hur många fall originalqueryn presterar bäst, lika och sämre som de ostrukturerade respektive strukturerade CLIR-queries jämför vi prestation av originalqueryn i förhållande till prestation av de ostrukturerade och strukturerade queries över de använda 24 topics. Resultatet redovisar vi i tabell 3 som baseras på data från tabell 2. Dessutom jämför vi prestation av de strukturerade och ostrukturerade queries. Förhållanden ”större än”, ”lika med” och ”mindre än” mellan två jämförande querytyper representeras av respektive tecken: >, =, <.

Prestation	*Antal topics*
original.> ostrukt.	18
original.=ostrukt.	0
original.<ostrukt.	6
original.> strukt.	12
original.=strukt.	0
original.<strukt.	12
strukt.>ostrukt.	18
strukt.=ostrukt.	2
strukt.<ostrukt.	4

Tabell 3. Prestation av de tre querytypernas över enskilda topics (Siffror anger antal topics för varje kategori)

Återvinningseffektivitet av olika querytyper varierar för varje enskilt topic vid olika DCV-nivåer. I tabell 4 har vi jämfört parvis prestationen av varje querytyp för olika topics för att se hur den fördelar sig över elva DCV-nivåer. För detta har vi jämfört precisionsvärde för given querytyp, ett topic och ett DCV-nivå. (Ursprungliga uppgifter redovisas i tabell 8, bilaga 2)

Vid DCV=5 har originalqueryn visat högre precisionsvärde i 8 topics jämfört med den ostrukturerade queryn. I 11 topics har originalqueryn visat lika precisionsvärde som den ostrukturerade queryn och den strukturerade queryn. I 5 topics har originalqueryn visat mindre precisionsvärde än den ostrukturerade queryn. Vid DCV-nivåer från 10 till 100

har originalqueryn visat högre precisionsvärde än den ostrukturerade queryn i mer än hälften av topics (antal topics varierar mellan 14 och 17).

Vid DCV=5 har den strukturerade queryn visat lika precisionsvärde som originalqueryn i 11 topics. Vid DCV=10, 20, 30 samt 80, 90, 100 presterar originalqueryn bättre i större antal topics (antal topics varierar mellan 10 och 12) än den strukturerade queryn.

Den strukturerade queryn har visat högre precisionsvärde i större antal topics än den ostrukturerade queryn vid alla använda DCV-nivåer. Vid DCV=5 och 10 är dock antal topics där den strukturerade queryn har presterat bättre något mindre, 9 respektive 11 topics, medan vid DCV= 20-100 den strukturerade qyeryn har visat bättre prestation i mer än hälften av topics, 13 till 18 topics.

Prestation/DCV	*5*	*10*	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*	*100*
original. > ostrukt.	8	14	16	16	16	15	15	15	16	17	14
original. = ostrukt.	11	3	3	3	3	4	1	3	3	3	4
original. < ostrukt.	5	7	5	5	5	5	8	6	5	4	6
original. > strukt.	8	10	11	11	7	7	7	7	11	12	11
original. = strukt.	11	8	6	6	5	3	5	6	3	2	2
original. < strukt.	5	6	7	7	12	14	12	11	10	10	11
strukt. > ostrukt.	9	11	16	14	13	16	13	15	17	17	18
strukt. = ostrukt.	8	8	7	7	7	5	4	5	4	4	3
strukt. < ostrukt.	7	5	1	3	4	3	7	4	3	3	3

Tabell 4. Prestation av de tre querytyperna vid elva DCV-nivåer över enskilda topics. (Siffror anger antal topics för varje kategori)

6.2 Analys

I det här avsnittet jämför vi skillnader i återvinningseffektivitet mellan tre olika querytyper. För detta har vi beräknat försämring av återvinningseffektiviteten av de två CLIR-queries i förhållande till återvinningseffektiviteten av originalqueryn och förbättring av återvinningseffektiviteten av den strukturerade queryn i förhållande till återvinningseffektiviteten av den ostrukturerade queryn. Skillnaden mellan ostrukturerade och strukturerade queries åskådliggör hur *syn*-baserad strukturering påverkar återvinningseffektivitet. Dessutom kommer vi att analysera skillnader i prestation ifråga om olika topics.

6.2.1 Precision vid olika DCV-nivåer

Skillnaden i precision mellan originalqueryn och den ostrukturerade CLIR-queryn varierar vid olika DCV-nivå och ligger mellan 0,125 och 0,043. Som det framgår av Diagram

1 och Tabell 1 ligger den största skillnaden mellan de två querytyperna i intervallet mellan DCV-nivåer 5-30 och varierar mellan 0,091 och 0,125. Vid DCV=5, som framgår av Tabell 1, har både den originalqueryn och den ostrukturerade CLIR-queryn fått det högsta genomsittliga precisionsvärdet på 0,442 respektive 0,333 och vid DCV=100 precisionsmedelvärde gick ned till 0,148 respektive 0,101 (Tabell 5). Resultatet tyder på att ord för ord queryöversättning utan struktur väsentligt försämrar återvinningseffektiviteten, t.ex. vid DCV=5 försämras den med 24,7 %, vid DCV=10 med 32,2 %. I genomsnitt försämras återvinningseffektiviteten med 30,4 % (Tabell 5). En förklaring till detta är att till följd av den lexikala flertydiheten vissa termer får större värde i den ostrukturerade queryn och systemet returnerar dokument där dessa termer förekommer oberoende av vilket sammanhang och betydelse de har i topicet. Följden av detta är att man får ett stort antal irrelevanta dokument. Dessutom kan relevanta dokument förekomma relativt sent i träfflistan.

DCV-nivå	*Original.*	*Ostrukt.*	Försämring
5	0,442	0,333	24,7 %
10	0,388	0,263	32,2 %
20	0,308	0,209	32,1 %
30	0,269	0,178	33,8 %
40	0,235	0,164	30,2 %
50	0,213	0,148	30,5 %
60	0,182	0,139	23,6 %
70	0,179	0,127	29,1 %
80	0,167	0,117	29,9 %
90	0,187	0,108	42,2 %
100	0,148	0,101	31,8 %
MV	0,247	0,172	30,4 %

Tabell 5. Försämringen i återvinningseffektiviteten av den ostrukturerade queryn jämfört med originalqueryn.

Skillnaden i återvinningseffektiviteten mellan den engelska originalqueryn och den strukturerade CLIR-queryn är relativt liten och utgör i genomsnitt 2,4 %. Vid DCV-nivåer 5-30 presterar originalqueryn bättre än den strukturerade queryn. Den största skillnaden ligger vid DCV=10 där prestationen av den strukturerade queryn visar en försämring med 9,8 % (Tabell 6). Från och med DCV=40 presterar den strukturerade CLIR-queryn bättre än originalqueryn med undantag av DCV=90. Negativa värden i tabell 6 visar en förbättring i precision av den strukturerade queryn.

DCV-nivå	*Original.*	*Strukt.*	Försämring
5	0,442	0,417	5,7 %
10	0,388	0,350	9,8 %
20	0,308	0,306	0,6 %
30	0,269	0,257	4,5 %
40	0,235	0,236	-0,4 %
50	0,213	0,220	-3,3 %

60	0,182	0,199	-9,3 %
70	0,179	0,183	-2,2 %
80	0,167	0,170	-1,8 %
90	0,187	0,160	14,4 %
100	0,148	0,152	-2,7 %
MV	0,247	0,241	2,4 %

Tabell 6. Försämringen i återvinningseffektivitet, den strukturerade queryn jämfört med originalqueryn.

Skillnaden i återvinningseffektiviteten mellan den ostrukturerade CLIR-queryn och den strukturerade CLIR-queryn utgör i genomsnitt 40,1 % (tabell 7). Högsta precisionsvärdet för båda querytyperna ligger vid DCV=5 och uppgår till 0,333 för den ostrukturerade respektive och 0,417 för den strukturerade queryn. Vid DCV-nivåer 5 och 10 förbättras effektiviteten med 25,2 % respektive 33,1 % medan vid DCV-nivåer 90 och 100 förbättras den med 48,1 % respektive 50,5 %. Detta tyder på att strukturingen av lexikonbaserade CLIR-query med *#syn*-operator minskar verkan av flertydighetsfaktorn och väsentligt kan förbättra återvinningseffektiviteten. *#syn*-operatorn tillåter att vikten mellan termerna fördelas på så sätt att lexikonets synonymer får lika vikt och viktningen av översättningstermer håller sig närmare till termvikterna i originalqueryn.

Det problem som finns kvar är att den strukturerade queryn visar mindre förbättring vid låga DCV-nivåer (DCV=5, 10), 25,2 % respektive 33,1 %, jämfört med den genomsnittliga förbättringen som ligger på 40,1 %. Vi antar att det kan bero på att även vid *syn*-strukturingen värdet av viktiga termer kan minskas en del genom jämnare viktfördelning bland synonymer. Detta problem kan undersökas i framtida studier.

DCV-nivå	***Strukt.***	***Ostrukt.***	***Förbättring***
5	0,417	0,333	25,2 %
10	0,350	0,263	33,1 %
20	0,306	0,209	46,4 %
30	0,257	0,178	44,4 %
40	0,236	0,164	43,9 %
50	0,220	0,148	48,6 %
60	0,199	0,139	43,2 %
70	0,183	0,127	44,1 %
80	0,170	0,117	45,3 %
90	0,160	0,108	48,1 %
100	0,152	0,101	50,5 %
MV	0,241	0,172	40,1 %

Tabell 7. Förbättring i återvinningseffektivitet av den strukturerade queryn jämfört med den ostrukturerade queryn.

6.2.2 Topic för topic analys

I analysen av enkilda topics tittar vi närmare på hur enskilda termer och deras översättningar påverkar återvinningseffektiviteten.

För vissa topics (26, 31, 39, 57) visar resultaten större skillnader mellan originalqueryn, de ostrukturerade och strukturerade CLIR-queries. T.ex., för topic 26 ligger genomsnittlig precision för originalqueryn på 0,363 för den ostrukturerade queryn på 0,031 och för den strukturerade queryn på 0,409 (tabell 2). Den stora skillnaden mellan originalqueryn och den ostrukturerade CLIR-queryn är kopplad till att de tre termer som finns i originalqueryn motsvars av så många som 14 termer i den ostrukturerade queryn. Det tyder på att vikten mellan termerna fördelas ojämnt. Termen *exploration* i originalqueryn motsvaras av tre översättningstermer i den ostrukturerade queryn medan termen *investigation* motsvaras av tio. I den strukturerade queryn motsvaras varje term i originalqueryn av bara en synonymgrupp, och därför fördelas vikten jämt för varje term och därför förbättras effektiviteten drastiskt.

Strukturеringen har betydande effekt i de fall där CLIR-queries har bland andra termer exakt samma termer som finns i originalqueryn. T.ex. i topic 31, där den strukturerade queryn visade precision 0,535 (Tabell 2) vilket närmar sig precision av originalqueryn 0,546 medan den ostrukturerade queryn visade mycket mindre precision på 0,246. I vissa fall får man inte samma termer efter översättningen, t.ex., i topic 34 termen ”encryption” i originalqueryn motsvaras i CLIR-queries av termen *@kodning* vilket ledde till väsentlig effektivitetsförsämring.

I två topics (42, 57) fungerade den ostrukturerade queryn bättre än båda originalqueryn och den strukturerade queryn. En möjlig förklaring till det är att i topic 57 termen #uw2(waste disposal) i originalqueryn motsvaras av två nära synonymer: #uw2(waste disposal) och #uw2(waste management) i CLIR-queries. Termen #uw2(waste management) bredder tydligen det semantiska fältet och kan förekomma i andra dokument än termen #uw2(waste disposal). I den ostrukturerade queryn har termen #uw2(waste management) fått större vikt än i den strukturerade queryn eftersom vid grupperingen vikten fördelas jämnt mellan termerna inom synonymgrupp.

Topic 42 utgör ett undantag i vår undersökning. *Mercy killing* är en typiskt engelsk gerundiumfras som består av en avledning *killing* (gerundium) och ett bestämningsord *mercy*. Huvudbetydelsen av denna fras kommer från huvudordet *killing*. Den frasen kan översättas med en svensk sammansättning ”barmhärtighetsmord”. Men denna sammansättning saknas i det använda svensk-engelska lexikonet. Därför har vi bestämt att betrakta denna fras som två separata ord vilket givetvis påverkade resultatet. Vid översättningen har vi fått sex termer. Men vi antar att bara en av de termerna, *murder*, kan förekomma i diskursen om ämnet. Termen *murder* finns inte i originalqueryn som inte har återvunnit några relevanta dokument. I den strukturerade queryn grupperas termen *murder* med två andra termer, *homicide* och *assassination*, som är synonymer till termen men inte berör ämnet. Därför har termen *murder* tydligen fått större vikt i den ostrukturerade queryn och

den ostrukturerade queryn har därmed fått ett bättre resultat än både originalqueryn och den strukturerade queryn.

6.3 Sammanfattande slutsatser

Resultatet av vår undersökning visar att återvinningseffektivitet av lexikonbaserade engelska ostrukturerade queries, vilka har svenska som ursprung, väsentligt skiljer sig från återvinningseffektivitet av engelska originalqueries. De ostrukturerade queries fick sämre medelvärde över alla topics vid samtliga använda DCV-nivåer. I genomsnitt har de ostukturerade queries nått 69 % återvinningseffektivitet av originalqueries (siffror baseras på Tabell 1).

Återvinningseffektivitet av engelska originalqueries och återvinningseffektivitet av lexikonbaserade engelska strukturerade queries, vilka har svenska som ursprung, skiljer sig marginellt. I genomsnitt har de strukturerade queries nått 97 % återvinningseffektivitet av originalqueries (siffror baseras på Tabell 1). Den största skillnaden ligger vid DCV= 5, 10, där försämring i återvinningseffektivitet utgör 5,7 % respektiv 9,8 % (Tabell 6).

Skillnaden i återviningseffektivitet mellan lexikonbaserade engelska ostrukturerade queries med svenska som ursprung och lexikonbaserade engelska strukturerade queries, också med svenska som ursprung, är väldigt tydlig. De ostrukturerade queries har nått i genomsnitt 71 % återvinningseffektivitet av de strukturerade queries. Detta tillåter oss påstå att querystrukturering är ett effektivt sätt att förbättra prestation av CLIR.

7. Jämförelse med tidigare CLIR-studier

Resultatet av Pirkolas (1998) och andras studier tyder på att det finns en väsentlig skillnad i återvinningseffektivitet mellan originalqueryn och den ostrukturerade queryn och mellan de ostrukturerade och strukturerade queries. Resultatet av vår studie stödjer deras forskningsresultat. Skillnaden mellan originalqueryn och den ostrukturerade queryn är väsentlig.

Eftersom vi har använt ett annat mått än Pirkola (1998), nämligen precision vid olika DCV-nivåer, kan vi inte jämföra exakta siffror med Pirkolas studie där det används Precision-Recall kurvor. Men vi kan jämföra vårt resultat med hans slutsatser i helhet. Pirkola har kommit fram till att med CLIR-query man kan nå samma återvinningseffektivitet som med enspråkig query genom att använda strukturering och båda allmänt och domänspecifikt lexikon. Vi testade querystruktur där queryn översatts med bara ett allmänt lexikon. Resultatet visade att effektiviteten av den strukturerade queryn avsevärt närmar sig effektiviteten av originalqueryn, och precisionen utgör ca 97 % precision av originalqueryn.

Pirkola et al. (2003) påpekade till och med att struktureringen väsentligt kan förbättra återvinningseffektiviteten av långa queries samt förbättra effektivitet av korta queries i lite minde grad. I den här studien innehåller originalqueries från 2 till 9 termer. Vi upptäckte att querystrukturering har stor effekt på queries som ursprungligen består av 3-5 termer t.ex., för topic 39 där originalqueryn består av 5 termer, utgör precision för den ostrukturerade queryn 0,026 och för den strukturerade queryn 0,137. I vår undersökning har vi bara två topics, 40 och 50, där query innehåller mer än 5 termer, 9 respektive 6 termer. Vi upptäckte att struktureringen hade betydande effekt. För topic 40 utgör precisionen för den ostrukturerade queryn 0,070 medan för den strukturerade queryn har en precision på 0,182, för topic 50 – 0,334 respektive 0,355.

Men i CLIR är längden av originalqueryn inte den enda faktor som påverkar effektiviteten av querystrukturering. Vissa termer, som högfrekventa ord, kan få fler översättningar och vissa av dem kan ha väldigt skilda betydelser i målspråket. Vilket i sin tur leder till försämring av återvinningseffektivitet hos CLIR-queries.

Hull (1997) antar att viktiga termer (lågfrekventa ord) ofta har 1-2 översättningsmotsvarigheter medan mindre viktiga termer har flera översättningsmotsvarigheter. Vår undersökning stödjer detta antagande, t.ex. i topic 31: "Identify documents discussing cases where rabies have been confirmed and what, if anything, is being done about it" har termen *rabies* bara två översättningsalternativ: *rabies hydrophobia* medan högfrekventa ordet *fall* har så många som 19 översättningsalternativ: *fall, backfall, tumble, descent, decline, drop, downfall, collapse, ruin, cadence, slope, gradient, declivity, rake, drop, case, event, instance, halyard.*

Ballesteros och Croft (1997) påstår att en enkel queryöversättning med ett lexikon kan försämra prestation med 40-60 %. De anser att en bra frasöversättning väsentligt kan förbättra resultatet och en dålig frasöversättning kan bidra till sämre resultat. I denna studie upptäckte vi att lexikonbaserade queryöversättning försämrade återvinningseffektivitet upp till 42 % (tabell 3). Vi har gjort en enkel ordbaserad queryöversättning. Vi översatte fraser som finns i lexikonet med sammansatta ord och de som inte finns i lexikonet ordagrant. I vår studie jämförde vi inte resultat av frasöversättning och ordagrann översättning. Därför kan inte förbättring med frasöversättning jämföras med deras forskningsresultat.

Resultatet av undersökningen som genomfördes av Hull och Grefenstette (1996) visade att genomsnittlig precision av de automatiskt översatta ordbaserade CLIR-queries uppnår ca 60 % av genomsnittlig precision av engelska originalqueries. (Precisionen beräknades vid DCV= 5, 10, 15 och 20.) Ordbaserade CLIR-queries formulerade med manuellt skapade tvåspråkig transfererbar ordlista uppnår 68 %. Resultatet av vår undersökning visar att man med querystrukturering kan nå mycket högre precision. Vid DCV= 5, 10, 20 har den strukturerade queryn nått i genomsnitt 90 % (94 %, 79 % respektive 99 %) av precisionen för originalqueryn (data baseras på Tabell 1).

8. Sammanfattning

Uppsatsen ägnas åt lexikonbaserad Cross-Language Information Retrival och innehåller översikt av för CLIR centrala problem och de existerande metoderna inom forskningsområdet. De viktigaste lingvistiska fenomen inom IR och CLIR har förklarats samt tidigare forskning har redovisats. Sedan har ett CLIR-experiment presenterats.

Syftet med undersökningen var att studera querystruktur i lexikonbaserad svensk-engelsk CLIR. Vi ställde följande forskningsfrågor:

- Hur skiljer sig återvinningseffektiviteten mellan engelska originalqueries och lexikonbaserade engelska ostrukturerade queries, vilka har svenska som ursprung?
- Hur skiljer sig återvinningseffektiviteten mellan engelska originalqueries och lexikonbaserade engelska strukturerade queries, vilka har svenska som ursprung?
- Hur skiljer sig återvinningseffektivitet mellan lexikonbaserade engelska ostrukturerade queries med svenska som ursprung och lexikonbaserade engelska strukturerade queries också med svenska som ursprung?
- Hur ser resultaten ut i förhållande till tidigare forskning?

Vi har utvärderat återvinningseffektiviteten hos tre olika typer av queries. För detta använde vi oss av 24 TREC-topics från vilka tre typer av queries formulerades. Samtliga queries kördes i det probabilistiska IR-systemet InQuery med hjälp av Query Performance Analyser. Vi undersökte skillnaderna i återvinningseffektivitet mellan originalqueries och två CLIR-queries. Undersökningen begränsades med DCV=100, som innebär att de 100 första dokumenten i återvinningsträfflistan har använts. Sammanlagt omfattade vår studie 2400 dokument för varje querytyp och som totalt för alla tre querytyper utgör 7200 dokument. De mått som vi använde i studien var precision vid elva DCV-nivåer: 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100. Medelvärden vid varje DCV-nivå för tre querytyper har jämförts. För att ge en bild av hur prestationen för varje querytyp skiljer sig för olika topics har vi jämfört parvis genomsnittlig precision för varje querytyp för varje topic.

Resultatet visade att återvinningseffektivitet skiljer sig väsentligt mellan originalqueries och ostrukturerade CLIR-queries. Skillnaden i återvinningseffektivitet mellan originalqueries och strukturerade CLIR-queries var liten. Skillnaden mellan de två typerna av CLIR-queries, ostrukturerade och strukturerade är också stor och tyder på att strukturereringen kan bidra till dramatisk förbättring av återvinningseffektiviteten. Vårt resultat stämmer med tidigare forskningsresultat.

Vi har kommit till följande slutsatser. En query som översätts med ett maskinläsbart lexikon, och där struktur inte användes i den resulterande queryn, har betydligt sämre återvinningseffektivitet än en enspråkig query. Effektivitetsförsämring orsakas åtminstone delvis av flertydighet som medförs vid automatisk översättning av querytermer. Strukturering av queries visade sig vara en effektiv metod för att hantera problem med översättningsflertydighet. På så sätt kan effektiviteten av CLIR-query betydligt förbättras och närma sig den återvinningseffektiviteten som man når med enspråkig IR.

Referenser

Adriani, Mirna (2001). English-Dutch CLIR using query translation techniques. Ingår i *Evaluation of Cross-Language Information Retrival Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF-2001*. Berlin: Springer Verlag, 2002. S.219-225. (Lecture Notes in Computer-Science, Vol. 2406.) Tillgängligt via http://www.ercim.org/publication/ws-proceedings/CLEF2/adriani.pdf. (05-04-26)

Ahlgren, Per (2004). *The effects of indexing strategy-query term combination on retrieval effectiveness in a Swedish full text database*. Borås: Valfrid (Publications from Valfrid, nr 28). Diss. University College of Borås/Göteborg university.

Ahlgren, Per, Borås. Personligt samtal 2005-05-20.

Ahlgren, Per & Eklund, Johan (2004). *Manual för Query Performance Analyser.* Borås: BHS: Kursmaterial för kusen ”Information Retrival”. Stencil.

Baeza-Yates, Ricardo & Ribeiro-Neto, Berthier (1999). *Modern Information Retrieval*. New York: ACM Press.

Ballesteros, Lisa & Croft, William Bruce (1996) Dictionary-based methods for cross-lingual information retrieval. Ingår i *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, 9-13 September, Zürich, Switzerland.* Roland Wagner, Helmut Thoma , eds. S.791-801.

Ballesteros, Lisa & Croft, William Bruce (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. Ingår i *Proceedings of the 20th annual international ACM SIGIR conference on Research and Development in Information Retrieval, 27-31 July, Philadelphia, Pennsylvania, USA.* Nicholas J.Belkin, Desai Narasimhalu, Peter Willett, eds. (Special issue of the SIGIR Forum, vol. 31.) New York: ACM Press. S.84 – 91.

Bolander, Maria (2001). *Funktionell svensk grammatik*. Stockholm: Liber.

Gey, Fredric, Jiang, Hailing & Chen, Aitao. (2003). Multilingual information retrieval. Ingår i *Encyclopedia of Library and Information Science.* 4 vol. Miriam A.Drake, ed. New York: Dekker. Vol.3. S.1895-1905.

Hedlund, Turid (2002) Compounds in dictionary-based cross-language information retrieval. *Information Reseach* vol.7, no.2. Tillgängligt via http://InformationR.net/ir/7-2/paper128.html

Hedlund, Turid (2003). *Dictionary-based cross-language information retrieval: principles, system design and evaluation*. Tampere: University of Tampere. (Acta Universitatis Tamperensis 962). Diss. Ubiversity of Tampere.

Hedlund, Turid et al. (2003). Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000-2002. *Information Retrieval* vol.7 no1/2. S.99-119.

Hull, David (1997). Using structured queries for disambiguation in cross-language information retrieval. Ingår i *Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval.* Stanford, CA. Tillgängligt även via URL: http://www.ee.umd.edu/medlab/filter/sss/papers/ (2005-03-05)

Hull, David (1998). A wieghted Boolean model for cross-language text retrieval. Ingår i *Cross-language Information Retrieval,* G.Grefenstette, ed. Boston: Kluwer Academic, S 119-136.

Hull, David & Grefenstette, Gregory (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. Ingår i *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 18-22 August, Zürich, Switzerland*. New York: ACM Press. S.49-57.

Ingo, Rune (1991). *Från källspåk till målspråk: Introduktion i översättningsvetenskap*. Lund: Studentlitteratur.

InQuery/TREC-Uta Search Guide. URL: http://www.info.uta.fi/qpaservlet/qpa5static/inquery_trec-uta_guide.html. (2005-04-20)

Kekäläinen, Jaana & Järvelin, Kalervo. (1998) The impact of query structure and query expansion on retrieval performance. Ingår i *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 24-28 August, Melbourne, Australia.* New York: ACM Press. S.130-137.

Moens, Maria-Francine (2000). *Automatic indexing and abstracting of document text*. Boston: Kluwer Academic.

Oard, Douglas W. & Diekema, Anne (1998). Cross language information retrieval. Ingår i *Annual review of information science and technology.* Martha E. Williams, ed. Medford, NJ: Information Today. Vol. 33. S.223–256.

Oard, Douglas W. & Dorr, Bonnie J. (1996). *A Survey of Multilingual Text Retrieval: Technical Report UMIACS-TR-96-19*. Maryland:University of Maryland, Institute for Advanced Computer Studies.

Pirkola, Ari (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. Ingår i *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, 24-28 August, Melbourne, Australia*. New York: ACM Press. S.55–63.

Pirkola, Ari (1999). *Studies on linguistic problems and methods in text retrieval : the effects of anaphor and ellipsis resolution in proximity searching, and translation and query structuring methods in cross-language retrieval.* Tampere: University of Tampere. (Acta Universitatis Tamperensis, 672) Diss. Ubiversity of Tampere.

Pirkola, Ari (2001). Morfolo gical typology of languages for IR. *Journal of Documentation* vol.57, nr.3. S.330-348. Tillgänglit via URL: www.info.uta.fi/tutkimus/fire/archive/morphological_**typology**.pdf (2005-04-27)

Pirkola, Ari, et al. (2001). Dictionary-based cross-language information retrieval: problems, methods, and research findings. *Information Retrieval* vol.4, no 3/4. S.209–230.

Pirkola, Ari, Puolamäki, Deniz & Järvelin, Kalervo (2003). Applying query structuring in cross-language retrieval. *Information Processing & Management*, vol. 39, iss.3. S.391-402.

Peters, Carol (2000). *CLEF - Cross-language evaluation forum*. Tillgängligt via: Cross-Language Evaluation Forum URL: http://clef.iei.pi.cnr.it/

Sheridan, Paraic & Smeaton Alan F. (1992). The Application of Morpho-Syntactical Language processing to effective phrase matching. *Information Processing & Management*, vol.28 iss.3. S.349-369.

Sormunen, Eero (2002). Liberal relevance criteria of TREC -: counting on negligible documents? *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 11-15 August, Tampere, Finland*. New York: ACM Press. S.324-330.

Sormunen, Eero, Halttunen, Kai & Keskustalo, Heikki (2002). *Query Performance Analyser - a tool for bridging information retrieval research and education.* Tampere: University of Tampere, Department of Information Studies, Research Notes 2002-1.

Strzalkowski, Tomek. (1995) Natural language information retrieval. *Information Processing & Management*, vol.31, iss.3. S.397-417.

Bilaga 1.

Topics, originalqueries, översättning till svenska samt ostrukturerade och strukturerade CLIR-queries

25. **Topic** 351: What information is available on petroleum exploration in the South Atlantic near the Falkland Islands?

Originalquery: #sum(@falkland island petroleum exploration)

Övers.: #sum((@falkland ö petroleum utforskning)

Osrtukt.: #sum(@falkland island isle petroleum #uw2(mineral oil) exploration #uw2(finding out) investigation)

Strukt.: #sum(@falkland #syn(island isle) #syn(petroleum #uw2(mineral oil)) #syn(exploration #uw2(finding out) investigation))

26. **Topic** 353: Identify systematic explorations and scientific investigations of Antarctica, current or planned.

Originalquery: #sum(exploration investigation @antarctica)

Övers.: #sum(utforskning undersökning @antarctica)

Osrtukt.: #sum(exploration #uw2(finding out) investigation examination inspection scrutiny search inquiry investigation poll study test testing @antarctica)

Strukt: #sum(#syn(exploration #uw2(finding out) investigation) #syn(examination inspection scrutiny search inquiry investigation poll study test testing) @antarctica)

28. **Topic 358**: What role does blood-alcohol level play in automobile accident fatalities?

Originalquery: #sum(blood alcohol level #uw2(automobile accident) fatality)

Övers.: #sum(blod alkohol halt bilolycka dödsolycka)

Osrtukt.: #sum(blood bloodstream alcohol content percentage substance worth halt lame limping #uw2(car accident) #uw2(fatal accident))

Strukt.: #sum(#syn(blood bloodstream) alcohol #syn(content percentage) #syn(substance worth) halt #syn(lame limping) #uw2(car accident) #uw2(fatal accident))

29. **Topic** 360: What are the benefits, if any, of drug legalization?

Originalquery: #sum(drug legalization benefit)

Övers.: #sum(drog legalisation fördel)

Osrtukt: #sum(drug legalization authentication advantage benefit profit advantage vantage van)

Strukt.: #sum(drug #syn(legalization authentication) #syn(advantage benefit profit advantage vantage van))

31. **Topic** 364: Identify documents discussing cases where rabies have been confirmed and what, if anything, is being done about it.

Originalquery: #sum(confirm rabies case)

Övers.: #sum(bekräfta rabies fall)

Osrtukt.: #sum(confirm corroborate #uw2(bear out) substantiate affirmsupport endorse acknowledge ratify certify rabies hydrophobia fall backfall tumble descent decline drop downfall collapse ruin cadence slope gradient declivity rake drop case event instance halyard)

Strukt.: #sum(#syn(confirm corroborate #uw2(bear out) substantiate affirmsupport endorse acknowledge ratify certify) #syn(rabies hydrophobia) #syn(fall backfall tumble descent decline drop downfall collapse ruin cadence slope gradient declivity rake drop) #syn(case event instance) halyard)

34. **Topic** 373: Identify documents that discuss the concerns of the United States regarding the export of encryption equipment.

Originalquery: #sum(@usa encryption equipment export)

Övers.: #sum(@usa kodning utrustning export)

Osrtukt.: #sum(@usa @kodning equipment outfit kit #uw2(fitting out) export exportation exports)

Strukt.: #sum(@usa @kodning #syn(equipment outfit) kit #uw2(fitting out) #syn(export exportation exports))

36. **Topic** 378: Identify documents that discuss opposition to the introduction of the euro, the European currency.

Originalquery: #sum(european currency euro opposition)

Övers.: #sum(europeisk valuta euro motstånd)

Osrtukt.: #sum(european #uw2(european community) currency exchange #uw2(foreign exchange) value euro resistance opposition surrender resistor rheostat)

Strukt.: #sum(#syn(european #uw2(european community)) #syn(currency exchange #uw2(foreign exchange)) value euro #syn(resistance opposition) #syn(surrender resistor rheostat))

39. **Topic** 387: Identify documents that discuss effective and safe ways to permanently handle long-lived radioactive wastes.

Originalquery: #sum(long-lived radioactive waste safe handle)

Övers.: #sum(långvarig radioaktiv avfall säker hantera)

Osrtukt.: #sum(long prolonged protracted lingering radioactive refuse rubbish waste #uw2(waste products) garbage trash offal #uw2(falling away) backsliding defection desertion apostasy sure certain positive confident safe secure reliable sure steady assured unerring infallible proof handle manage wield use #uw2(make use of) treat restrain check)

Strukt.: #sum(#syn(long prolonged protracted lingering) radioactive #syn(refuse rubbish waste #uw2(waste products) garbage trash offal) #syn(#uw2(falling away) backsliding defection desertion apostsy) sure certain positive confident safe secure reliable sure steady assured unerring infallible proof #syn(handle manage wield use #uw2(make use of) treat restrain check))

40. **Topic** 388: Identify documents that discuss the use of organic fertilizers (composted sludge, ash, vegetable waste, microorganisms, etc.) as soil enhancers.

Originalquery: #sum(organic fertilizer sludge ash vegetable waste microorganism soil enhancer)

Övers.: #sum(organisk gödningsmedel slam aska grönsak avfall mikroorganism jord förbättrare)

Ostrukt.: #sum(organic fertilizer slam ashes ash vegetable refuse rubbish waste #uw2(waste products) garbage garbage trash offal #uw2(falling away) backsliding defection desertion apostasy micro-organism ground soil dirt earth

dust land earth world earth ground improve amend reform ameliorate #uw2(improve upon))

Strukt.: #sum(organic fertilizer slam #syn(ashes ash) vegetable #syn(refuse rubbish waste #uw2(waste products) garbage) #syn(garbage trash offal) #syn(#uw2(falling away) backsliding) #syn(defection desertion) apostasy micro-organism #syn(ground soil dirt earth dust) land #syn(earth world) #syn(earth ground) #syn(improve amend reform ameliorate #uw2(improve upon)))

42. **Topic** 393: Identify documents that discuss mercy killings.

Originalquery: #sum(mercy killing)

Övers.: #sum(barmhärtighet mord)

Ostrukt.: #sum(mercy compassion charity murder homicide assassination)
Strukt.: #sum(#syn(mercy compassion charity) #syn(murder homicide assassination))

43. **Topic** 396: Identify documents that discuss sick building syndrome or building-related illnesses.

Originalquery: #sum(sick building syndrome illness)

Övers.: #sum(sjuk byggnad syndrom sjukdom)

Ostrukt.: #sum(sick unwell invalid ailing bad diseased disordered morbid suspicious shady fishy building edifice build structure syndrome illness ill-health disease malady disorder complaint ailment affection sickness)

Strukt.: #sum(#syn(sick unwell invalid ailing bad diseased disordered) #syn(morbid suspicious shady fishy) #syn(building edifice build structure) syndrome #syn(illness ill-health disease malady disorder complaint ailment affection sickness))

45. **Topic** 400: What measures are being taken by local South American authorities to preserve the Amazon tropical rain forest?

Originalquery: #sum(amazon #uw2(rain forest) preserve #uw2(south american) authority)

Övers.: #sum(amason regnskog skydda sydamerikan myndighet)

Ostrukt.: #sum(amazon #uw2(rain forest) protect shelter shield defend cover guard safeguard preserve secure #uw2(south american) authority #uw2(public authority) authority power authoritativeness majority #uw2(full age))

Strukt.: #sum(amazon #uw2(rain forest) #syn(protect shelter shield defend cover guard safeguard preserve secure) #uw2(south american) #syn(authority #uw2(public authority)) #syn(authority power) authoritativeness #syn(majority #uw2(full age)))

46. **Topic** 402: What is happening in the field of behavioral genetics, the study of the relative influence of genetic and environmental factors on an individual's behavior or personality?

Originalquery: #sum(behavioral genetics genetic environmental factor influence personality behavior)

Övers.: #sum(beteende genetik genetisk miljö faktor inflytande personlighet beteende)

Ostrukt.: #sum(behaviour conduct genetics genetic environment milieu surroundings setting factor element foreman overseer influence ascendancy sway effect personality personage personage figure behaviour conduct)

Strukt.: #sum(#syn(behaviour conduct) genetics genetic #syn(environment milieu surroundings setting) #syn(factor element) #syn(foreman overseer) #syn(influence ascendancy sway effect) personality #syn(personage figure) #syn(behaviour conduct))

47. **Topic** 403: Find information on the effects of the dietary intakes of potassium, magnesium and fruits and vegetables as determinants of bone mineral density in elderly men and women thus preventing osteoporosis (bone decay).

Originalquery: #sum(dietary intake potassium magnesium fruit vegetable prevent @osteoporosis)

Övers.: #sum(dietisk intag potassium magnesium frukt grönsak förhindra @osteoporosis)

Ostrukt.: #sum(dietary dietetic intake inlet potato spud potatoes magnesium fruit fruits product result outcome issue vegetable greens prevent @osteoporosis)

Strukt.: #sum(#syn(dietary dietetic) #syn(intake inlet) #syn(potato spud potatoes) magnesium #syn(fruit fruits product result outcome issue) #syn(vegetable greens) prevent @osteoporosis)

48. **Topic** 405: What unexpected or unexplained cosmic events or celestial phenomena, such as radiation and supernova outbursts or new comets, have been detected?

Originalquery: #sum(unexpected unexplained cosmic event celestial phenomenon radiation supernova outburst comet)

Övers.: #sum (oväntad oförklarad kosmisk händelse celest fenomen strålning supernova utbrott komet)

Ostrukt.: #sum (unexpected, #uw2(unlooked for) oförklarad cosmic phenomenon fact celestial phenomenon radiation supernova outbreak eruption outburst burst fit explosion ebullience comet)

Strukt.: #sum (#syn(unexpected, #uw2(unlooked for)) oförklarad cosmic #syn(phenomenon fact) celestial phenomenon radiation supernova #syn(outbreak eruption outburst burst fit explosion ebullience) comet)

49. **Topic** 407: What is the impact of poaching on the world's various wildlife preserves?

Originalquery: #sum(poaching wildlife preserve)

Övers.: #sum(tjuvskytte vild djur reservat)

Ostrukt.: #sum(poaching #uw2(game poaching) wild savage uncivilized untamed unruly furious animal beast insect thing reserve, #uw2(national park) #uw2(national reserve park) sanctuary #uw2(game preserve) #uw2(wild life preserve) reservation)

Strukt.: #sum(#syn(poaching #uw2(game poaching)) #syn(wild savage uncivilized untamed unruly furious) #syn(animal beast insect thing) #syn(reserve, #uw2(national park) #uw2(national reserve park) sanctuary #uw2(game preserve) #uw2(wild life preserve) reservation))

50. **Topic** 408: What tropical storms (hurricanes and typhoons) have caused significant property damage and loss of life?

Originalquery: #sum(tropical storm hurricane typhoon property damage life loss)

Övers.: #sum(tropisk storm orkan tyfon egendom skadegörelse liv förlust)

Ostrukt.: #sum(tropical tropic gale storm tempest storm assault topper hurricane typhoon siren property estate property damage life lifetime existence vitality #uw(way of life) living #uw2(living being) soul #uw2(living soul) body waist bodice row noise commotion to-do fuss loss damage forfeiture)

Strukt.: #sum(#syn(tropical tropic) #syn(gale storm tempest) #syn(storm assault) topper hurricane typhoon siren property estate property damage #syn(life lifetime existence vitality #uw2(way of life) living) #syn(#uw2(living being) soul #uw2(living soul)) body waist bodice #syn(row noise commotion to-do fuss) #syn(loss damage forfeiture))

56. **Topic** 420: How widespread is carbon monoxide poisoning on a global scale?

Originalquery: #sum(#uw2(carbon monoxide) poison)

Övers.: #sum(koloxid gift)

Ostrukt.: #sum(#uw2(carbon monoxide) poison venom virus toxin married)

Strukt.: #sum(#uw2(carbon monoxide) #syn(poison venom virus toxin) married)

57. **Topic** 421: How is the disposal of industrial waste being accomplished by industrial management throughout the world?

Originalquery: #sum(world industrial management #uw2(waste disposal))

Övers.: #sum(värld industriell hantering avfallshantering))

Ostrukt.: #sum(world earth industrial handling trade business #uw2(waste disposal) #uw2(waste management))

Strukt.: #sum(#syn(world earth) industrial #syn(handling trade business) #syn(#uw2(waste disposal) #uw2(waste management)))

58. **Topic** 427: Find documents that discuss the damage ultraviolet (UV) light from the sun can do to eyes.

Originalquery: #sum(ultraviolet eye damage)
Övers.: #sum(ultraviolett öga skada)

Ostrukt.: #sum(ultraviolet eye pip eyelet loop injury damage lesion harm mischief loss detriment disadvantage injure hurt #uw2(be bad for) #uw2(be detrimental to) prejudice #uw2(do harm to) impair)

Strukt.: #sum(ultraviolet eye pip eyelet loop #syn(injury damage lesion harm mischief loss detriment disadvantage) #syn(injure hurt damage #uw2(be bad for) #uw2(be detrimental to) prejudice #uw2(do harm to) impair))

61. **Topic** 437: What has been the experience of residential utility customers following deregulation of gas and electric?

Originalquery: #sum(customer experience gas electric deregulation)

Övers.: #sum(kund uppleva gas el avreglering)

Ostrukt.: #sum(customer patron client experience know #uw2(meet with) #uw2(take part in) #uw2(live through) #uw2(go through) witness see spend feel #uw2(react to) gas electricity deregulation)

Strukt.: #sum(#syn(customer patron client) #syn(experience know #uw2(meet with) #uw2(take part in) #uw2(live through) #uw2(go through) witness see spend feel #uw2(react to)) gas electricity deregulation)

62. **Topic** 440: What steps are being taken by governments or corporations to eliminate abuse of child labor?

Originalquery: #sum(eliminate #uw2(child labor))

Övers.: #sum (eliminera barnarbete)

Ostrukt.: #sum(eliminate #uw2(child labour) #uw2(employment of children) #uw2(employment of children and young persons))

Strukt.: #sum(eliminate #syn(#uw2(child labour) #uw2(employment of children) #uw2(employment of children and young persons)))

64. **Topic** 445: What other countries besides the United States are considering or have approved women as clergy persons?

Originalquery: #sum(approval woman clergy)

Övers.: #sum(godkännande kvinnlig prästerskap)

Ostrukt.: #sum(approving approbation approval confirmation admittance acknowledgement acceptance female woman feminine womanly women's ladies' womanish effeminate clergy clergymen priesthood priests)

Strukt.: #sum(#syn(approving approbation approval confirmation admittance acknowledgement acceptance) #syn(female woman feminine womanly women's ladies' womanish effeminate) #syn(clergy clergymen priesthood priests))

65. **Topic** 448: Identify instances in which weather was a main or contributing factor in the loss of a ship at sea.

Originalquery: #sum(weather sea ship loss)

Övers.: #sum(väder sjö fartyg förlust)

Ostrukt.: #sum(weather air wind lake sea wave pool vessel ship craft loss damage forfeiture)

Strukt.: #sum (weather #syn (air wind) #syn(lake sea wave pool) #syn(vessel ship craft) #syn(loss damage forfeiture))

Bilaga 2

Precision för enskilda queries vid olika DCVer

To-pic	Querytyp/ DCV	5	10	20	30	40	50	60	70	80	90	100
25	Original.	0,60	0,70	0,60	0,43	0,35	0,30	0,25	0,21	0,19	0,17	0,16
	Otrukt.	0,60	0,60	0,55	0,40	0,30	0,24	0,22	0,19	0,16	0,14	0,13
	Strukt.	0,60	0,60	0,55	0,40	0,38	0,30	0,27	0,23	0,21	0,19	0,17
26	Original.	0,80	0,50	0,40	0,40	0,35	0,36	0,30	0,26	0,23	0,21	0,19
	Otrukt.	0,00	0,00	0,05	0,03	0,03	0,04	0,05	0,04	0,04	0,03	0,03
	Strukt.	0,80	0,60	0,45	0,47	0,43	0,40	0,35	0,30	0,26	0,23	0,21
28	Original.	0,20	0,10	0,20	0,13	0,10	0,10	0,10	0,09	0,09	0,09	0,10
	Otrukt.	0,80	0,40	0,30	0,23	0,20	0,16	0,15	0,16	0,16	0,14	0,13
	Strukt.	0,60	0,50	0,30	0,27	0,20	0,22	0,22	0,19	0,16	0,16	0,15
29	Original.	0,80	0,80	0,85	0,87	0,88	0,82	0,78	0,77	0,76	0,76	0,74
	Otrukt.	1,00	0,90	0,90	0,83	0,73	0,70	0,65	0,60	0,58	0,52	0,47
	Strukt.	0,80	0,80	0,85	0,90	0,88	0,88	0,80	0,80	0,76	0,74	0,73
31	Original.	0,80	0,90	0,65	0,57	0,53	0,52	0,52	0,46	0,40	0,36	0,32
	Otrukt.	0,80	0,50	0,30	0,20	0,15	0,14	0,15	0,13	0,11	0,11	0,11
	Strukt.	0,80	0,70	0,75	0,53	0,53	0,58	0,50	0,44	0,39	0,34	0,32
34	Original.	0,60	0,50	0,25	0,17	0,13	0,10	0,08	0,07	0,06	0,07	0,06
	Otrukt.	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Strukt.	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
36	Original.	0,00	0,10	0,10	0,13	0,10	0,08	0,07	0,06	0,05	0,04	0,04
	Otrukt.	0,00	0,00	0,00	0,03	0,05	0,04	0,03	0,04	0,04	0,03	0,03
	Strukt.	0,00	0,00	0,05	0,03	0,03	0,02	0,03	0,03	0,03	0,02	0,02
39	Original.	0,20	0,30	0,25	0,20	0,15	0,16	0,17	0,14	0,15	0,14	0,13
	Otrukt.	0,00	0,00	0,00	0,00	0,03	0,04	0,03	0,04	0,04	0,04	0,06
	Strukt.	0,00	0,10	0,20	0,13	0,18	0,18	0,17	0,14	0,14	0,14	0,13
40	Original.	0,40	0,30	0,30	0,27	0,20	0,18	0,17	0,14	0,14	0,12	0,11
	Otrukt.	0,00	0,10	0,10	0,10	0,10	0,08	0,07	0,06	0,06	0,06	0,05
	Strukt.	0,60	0,40	0,20	0,13	0,10	0,10	0,08	0,10	0,10	0,10	0,09
42	Original.	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Otrukt.	0,20	0,10	0,05	0,03	0,03	0,02	0,02	0,01	0,01	0,01	0,01
	Strukt.	0,00	0,00	0,05	0,03	0,03	0,02	0,02	0,01	0,03	0,02	0,02
43	Original.	1,00	1,00	0,55	0,37	0,28	0,22	0,18	0,16	0,14	0,12	0,11
	Otrukt.	0,80	0,70	0,45	0,40	0,30	0,26	0,25	0,21	0,19	0,17	0,16
	Strukt.	1,00	0,80	0,50	0,37	0,33	0,26	0,22	0,20	0,19	0,18	0,16
45	Original.	0,80	0,60	0,60	0,60	0,56	0,48	0,48	0,44	0,40	0,39	0,35
	Otrukt.	0,60	0,70	0,56	0,50	0,43	0,38	0,35	0,30	0,28	0,24	0,22
	Strukt.	0,60	0,60	0,60	0,53	0,50	0,46	0,43	0,40	0,38	0,36	0,34
46	Original.	0,40	0,40	0,25	0,27	0,23	0,18	0,15	0,13	0,14	0,13	0,12
	Otrukt.	0,40	0,20	0,15	0,13	0,13	0,10	0,08	0,07	0,08	0,07	0,06
	Strukt.	0,20	0,30	0,20	0,20	0,20	0,18	0,17	0,14	0,13	0,11	0,10
47	Original.	0,00	0,00	0,15	0,17	0,13	0,10	0,10	0,09	0,08	0,07	0,06
	Otrukt.	0,00	0,00	0,20	0,17	0,13	0,10	0,08	0,07	0,06	0,07	0,06

	Strukt.	0,40	0,40	0,25	0,17	0,13	0,12	0,10	0,09	0,09	0,08	0,07
48	Original.	0,40	0,20	0,20	0,20	0,18	0,16	0,15	0,13	0,11	0,12	0,12
	Otrukt.	0,40	0,30	0,15	0,20	0,15	0,14	0,15	0,13	0,11	0,11	0,11
	Strukt.	0,20	0,30	0,30	0,20	0,20	0,18	0,15	0,14	0,15	0,14	0,14
49	Original.	0,00	0,10	0,05	0,03	0,03	0,04	0,03	0,03	0,03	0,02	0,02
	Otrukt.	0,00	0,00	0,05	0,03	0,03	0,02	0,02	0,01	0,01	0,01	0,01
	Strukt.	0,20	0,10	0,10	0,07	0,05	0,08	0,07	0,07	0,06	0,06	0,05
50	Original.	0,40	0,30	0,35	0,37	0,40	0,38	0,03	0,34	0,31	0,30	0,31
	Otrukt.	0,60	0,50	0,25	0,03	0,38	0,38	0,33	0,34	0,31	0,28	0,27
	Strukt.	0,40	0,30	0,40	0,37	0,35	0,34	0,38	0,34	0,34	0,33	0,35
56	Original.	0,80	0,80	0,50	0,37	0,30	0,24	0,20	0,19	0,19	0,17	0,15
	Otrukt.	0,20	0,10	0,15	0,17	0,15	0,16	0,15	0,13	0,13	0,11	0,10
	Strukt.	0,60	0,70	0,50	0,37	0,30	0,24	0,20	0,19	0,16	0,14	0,13
57	Original.	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,02
	Otrukt.	0,00	0,00	0,00	0,03	0,03	0,02	0,03	0,03	0,03	0,02	0,02
	Strukt.	0,00	0,00	0,00	0,00	0,03	0,02	0,02	0,01	0,01	0,01	0,01
58	Original.	0,60	0,30	0,30	0,27	0,20	0,20	0,17	0,14	0,13	0,11	0,10
	Otrukt.	0,00	0,20	0,10	0,07	0,05	0,04	0,03	0,03	0,03	0,02	0,02
	Strukt.	0,60	0,30	0,25	0,20	0,18	0,14	0,12	0,10	0,09	0,09	0,09
61	Original.	0,00	0,10	0,05	0,07	0,05	0,04	0,03	0,04	0,05	0,04	0,04
	Otrukt.	0,00	0,10	0,05	0,03	0,05	0,04	0,05	0,04	0,04	0,04	0,05
	Strukt.	0,20	0,10	0,05	0,03	0,03	0,02	0,03	0,04	0,05	0,04	0,06
62	Original.	0,80	0,40	0,25	0,23	0,25	0,20	0,20	0,19	0,16	0,14	0,13
	Otrukt.	0,80	0,50	0,40	0,40	0,35	0,28	0,23	0,20	0,18	0,17	0,15
	Strukt.	0,80	0,50	0,40	0,40	0,35	0,28	0,23	0,20	0,18	0,17	0,15
64	Original.	0,60	0,70	0,40	0,27	0,20	0,18	0,15	0,16	0,14	0,12	0,11
	Otrukt.	0,60	0,30	0,20	0,17	0,13	0,12	0,13	0,13	0,11	0,11	0,11
	Strukt.	0,40	0,20	0,25	0,20	0,20	0,16	0,13	0,11	0,10	0,09	0,08
65	Original.	0,40	0,20	0,15	0,10	0,08	0,06	0,05	0,06	0,08	0,78	0,07
	Otrukt.	0,20	0,10	0,05	0,07	0,05	0,06	0,07	0,07	0,08	0,08	0,07
	Strukt.	0,20	0,10	0,15	0,17	0,13	0,10	0,10	0,10	0,09	0,08	0,08

Tabell 8. Precisionsvärde för ens kilda topics vid elva DCV-nivåer