Kandidatuppsats inom Kognitionsvetenskap 2012

En manuell utvärdering av extraktionssammanfattade texter.

Institutionen för datavetenskap
Författare: Marcus Johansson
Handledare: Arne Jönsson

Linköpings universitet

Sammanfattning

Att göra sammanfattningar med hjälp av datorer istället för att manuellt skriva dem är något som många försöker komma på en bra lösning för. Hur man ska utvärdera om de sammanfattningar som produceras är bra är dock ingen självklarhet och flera olika lösningar har föreslagits. I denna studie används det svenska högskoleprovet som grund för att undersöka hur stor skillnad det är mellan originaltexterna och deras motsvarande extraktionsbaserade sammanfattningar. I studien visas att det går nästan dubbelt så snabbt att läsa en trettioprocentig extraktionsbaserad sammanfattning jämfört med dess original. Skillnaden i antal rätt svar mellan sammanfattningarna och originaltexterna blev cirka 10 procent sämre för sammanfattningarna än originaltexterna. Sammanfattningarna skattades också som sämre i de subjektiva skattningar försöksdeltagarna fick göra för varje text.

Innehåll

Inledning 1

Avgränsningar. 1

Frågeställningar. 2

Bakgrund 3

Metod 6

Försöksdeltagare 6

Utformning av frågeformulär. 6

Pilottest. 7

Förändringar i frågeformulär ett. 7

Förändringar i frågeformulär två. 7

Testet 8

Resultat 10

Skillnader i antal rätt svar. 10

Skillnad i tid det tar att läsa. 10

Resultat från de subjektiva skattningsfrågorna. 11

Resultatdiskussion 13

Metoddiskussion 14

Slutsats 14

Referenser 17

Appendix 18

Inledning

Mycket av den information vi tar till oss i dagens samhälle är textbaserad. På internet finns det enorma mängder text om alla möjliga olika ämnen. För att inte behöva läsa all text för att bestämma om texten är av intresse för ens vidare förståelse av ämnet är det ofta fördelaktigt med någon form av sammanfattning. Många texter på internet saknar dock just en sammanfattning (Hassel 2004). Personer som har svårt för att läsa av olika anledningar, som synnedsättning eller dyslexi, eller personer med ett annat modersmål, kan också ha fördel av att kunna läsa en kortare version av olika texter (Smith & Jönsson 2011 A).

Olika distinktioner kan göras mellan sammanfattningar beroende på deras stil och syften. Tre huvudsakliga grupper kan identifieras. Dels kan sammanfattningar vara indikativa, alltså ge en bild av textens informationsinnehåll och ge en övergripande bild av ämnet. De kan också användas istället för originaltexten, och är då tänkta att innehålla all relevant information som finns i originaltexten. Den tredje gruppen är sammanfattningar som används för att kritisera originaldokumentet. (Mani et. al. 1999)

En annan distinktion som är vanlig är den mellan extraktionsbaserade sammanfattningar och abstracts (Mani et. al. 1999, Smith & Jönsson 2011 A). Skillnaden mellan extraktioner och abstracts är att ett abstract är en omskrivning av originaltexten, där den viktigaste informationen finns kvar men textmassan har reducerats kraftigt. En extraktionsbaserad sammanfattning är en sammanfattning där de viktigaste meningarna eller orden plockas ut ur ett dokument och ämnar till att plocka ut de meningar eller ord som ger bäst täckning av det ämne originaltexten tar upp. (Mani et. al. 1999, Smith & Jönsson 2011 A)

För att mäta hur komplex en text är används en rad olika mått, där LIX och OVIX är två av de vanligaste i Sverige. Vid Linköpings universitet har studier visat att extraktionsbaserade sammanfattningar ger ett lägre LIX och OVIX mått än många originaltexter, vilket betyder att texterna bland annat innehåller färre långa, och därmed komplexa, ord. Framförallt för populärvetenskapliga artiklar och nyhetsartiklar verkar extraktionsbaserade sammanfattningar göra texterna lättare att läsa (Smith & Jönsson 2011 B).

CogSum är ett program som är utvecklat vid Linköpings universitet vars syfte är att göra extraktionsbaserade sammanfattningar. Det använder sig bland annat av Random indexing och Page Rank (Smith & Jönsson 2011 A), en mer genomgående förklaring finns under kapitlet bakgrund.

Avgränsningar.

Studien avgränsades till att undersöka extraktionsbaserade sammanfattningar gjorda av CogSum, men ingen vikt lades vid själva programmet utan enbart de sammanfattningar som genererades av programmet. De flesta extraktionsbaserade sammanfattare arbetar med att plocka ut de viktigaste meningarna ur en text med liknande formler även om de skiljer sig på vissa punkter ((Chatterjee & Mohan 2007, Mani et al. 1999)) CogSum är förmodligen inte representativt för alla extraktionsbaserade sammanfattare, men de som arbetar på ett liknande sätt kommer förmodligen få ett liknande resultat. Det som testades var alltså inte programmets utformning, utan extraktionsbaserade sammanfattningar och hur de fungerar för att bevara information och hur sammanfattningarna upplevs av dem som läser texterna.

Frågeställningar.

För studien var ett par frågeställningar intressanta. Det var för studien ett mål att se hur bra extraktionsbaserade sammanfattningar fungerar för att ge en bra uppfattning om det ämne originaldokumentet handlar om. Därför var det intressant att undersöka om personer skulle kunna prestera lika bra med hjälp av sammanfattningarna som med originaldokumenten på ett läsförståelsetest.

Som ett komplement till prestation på testet är det också intressant att undersöka om det går att spara tid genom att läsa sammanfattningarna gentemot originaltexterna.

Det är också av intresse att undersöka hur extraktionsbaserade sammanfattningar upplevs av dem som läser dessa texter.

Bakgrund

Många har arbetat med att göra extraktionsbaserade sammanfattare så bra som möjligt. Flera olika metoder har föreslagits för att plocka ut den viktigaste informationen för att skapa extraktionsbaserade sammanfattningar. Enligt Chatterjee & Mohan bygger många av dessa på olika sätt att ge olika meningar poäng. Några olika sätt att skapa extraktionsbaserade sammanfattningar är med hjälp av ”Mutual Reinforcement Principle for Summary generation” som använder grupper av meningar för att ge dem olika poäng efter hur nära de ligger textens centrala tema (Chatterjee & Mohan 2007). Meningarna kan också få poäng efter olika fördefinierade egenskaper, som lingvistiska egenskaper och statistiska egenskaper som plats och struktur (Chatterjee & Mohan 2007).

Många extraktionsbaserade sammanfattare, däribland CogSum (Smith & Jönsson 2011 A), använder sig av en ordrymdsmodell som kallas random indexing. Det kan beskrivas som en tvåstegsmodell där en högdimensionell indexvektor, typiskt på flera hundra dimensioner, tilldelas varje kontext (ett ord eller ett dokument). Dessa vektorer består sedan av ett litet antal +1 och -1 och ett stort antal nollor, exempelvis fyra +1 och fyra -1 tillsammans med 800 nollor. Därefter skapas kontextvektorer genom att skanna igenom texten efter varje gång ett visst ord används, och orden runt detta ord är det som är kontexten, och därefter läggs de ordens indexvektorer till det andra ordets kontextvektor. Kontextvektorn för ett ord är alltså de båda orden före och efter ordets indexvektorer. På så vis får man fram hur ofta ett ord används i liknande kontexter. Ord ses alltså som kontextvektorer som är summan av ordets alla kontextvektorer (Sahlgren 2005, Smith & Jönsson 2011 A).

Tillsammans med random indexing använder CogSum något som kallas page rank, en algoritm som används för att ytterligare bestämma vilka meningar som är viktiga för texten. Det görs genom att med hjälp av vektorerna från random indexing beräkna vinklarna för de olika meningarna, mot varandra. Varje gång det görs förändras vinklarna för de andra meningarna något och så görs beräkningen om, för alla meningar i hela dokumentet mot varandra. Detta görs flera gånger tills förändringarna stabiliseras och de viktigaste meningarna kan plockas ut (Smith & Jönsson 2011 A).

Hur man ska utvärdera hur bra en sammanfattning fungerar är inte en helt självklar process. Många olika sätt har föreslagits, och två övergripande kategorier har vuxit fram. Dels talar man om en yttre (extrinsic) och en inre (intrinsic) utvärdering (Mani et. al. 1999, Hassel 2004). En inre utvärdering bygger på att man mäter systemet mot någon form av norm för vad som är en bra sammanfattning. (Hassel 2004) Det sker ofta mot en så kallad guldstandard, en sammanfattning som är gjord på liknande sätt som av maskinen, fast av människor (Carlsson 2009). Man mäter sedan antal överlappande meningar mellan guldstandard och maskinsammanfattning för att avgöra om sammanfattningen är bra eller inte. Ett problem med guldstandarder är dock att människor sällan är överens om vilka meningar som borde vara med i sammanfattningen och inte (Carlsson 2009). Yttre utvärderingar å andra sidan försöker mäta upp hur bra texten är för att fullfölja någon form av uppgift, och undersöka hur bra resultatet blir efter användning av sammanfattningen. Det finns flera olika sätt att göra både yttre och inre utvärderingar, en mer övergripande genomgång finns i (Hassel 2007).

Ett stort försök att utvärdera flera olika sammanfattare är ”the tipster summac text summarization evaluation”. I det testet var flera olika universitet delaktiga i att göra sammanfattningar för att utvärderas på ett gemensamt sätt. I studien användes tre olika utvärderingsmetoder, två yttre och en inre. Dels använde de sig av ett ad-hoc test, där ett dokument, sammanfattning eller originaltext,

skulle poängsättas tillsammans med sin rubrik efter sin relevans för ett visst ämne. Det visade sig att sammanfattningar på 17 procent fungerar likvärdigt som en originaltext för att kunna avgöra om informationen är relevant för ämnet, medan beslutsfattandetiden nästan halveras. De hade även ett test som de kallade för ”the categorization task” där generiska sammanfattningar skulle bedömmas efter om de innehöll tillräckligt med information för att personer skulle kunna kategorisera dem under rätt ämne. Här fick personerna fem olika ämnen, och skulle säga till vilken kategori texten tillhörde, eller om den inte passade in för någon av dem. Redan tio procentiga sammanfattningar presterade lika bra som sina motsvarande originaltexter, medan det gick nästan 40 procent snabbare att kategorisera sammanfattningarna (Mani et al. 1999).

De hade även en inre utvärderingsmetod, ett ”question and answering task”. För detta test fanns ett färdigt schema med olika meningar som räknades som svar på ett par olika frågor, sammanfattningarna jämfördes sedan med detta schema för att se om de tog ut rätt meningar. I detta test presterade sammanfattningar på mellan 30-40 procent bäst (Mani et al. 1999).

Chin-Yew & Hovy gjorde 2003 ett experiment för att manuellt kunna avgöra precision and recall för ett system, eftersom detta ofta görs med enbart en mening som rätt, medan flera meningar i en text kan innehålla ungefär samma information. Här skulle olika personer säga om en mening från en guldstandard och från en maskinsammanfattning överlappade varandra med information i fem olika steg; allt, nästan allt, en del, nästan ingenting och inget alls. Det visade sig dock vara ett svårt experiment och jämförelser mellan samma två meningar kunde få två, eller ibland tre olika kategoriseringar och ansågs därför inte som ett speciellt bra sätt att utvärdera sammanfattare på. (Chin-Yew & Hovy 2003)

De gjorde också en maskinell utvärdering med måttet BLEU, som jämför mängden överlappande n-gram mellan en sammanfattning och en guldstandard. Här kom de fram till att det krävs mer än en guldstandard för att kunna göra dessa utvärderingar och att samma sammanfattning bör utvärderas flera gånger för att få ett tillförlitligt resultat (Chin-Yew & Hovy 2003).

Jing, McKeown, Barzilay och Elhadad gjorde 1998 en utvärdering mellan sammanfattningar på 10 och 20 procent mot en guldstandard på samma längd. Guldstandarden gjordes av flera olika människor och visade att olika personer ofta väljer olika meningar som de viktigaste för en sammanfattning. De visade också att desto längre en sammanfattning ska vara, desto mer skiljer det sig åt mellan olika personer vilka meningar som ska ses som mest relevanta för ämnet. De ansåg att precision and recall inte var ett så bra sätt att utvärdera sammanfattningar på eftersom olika meningar kan vara lika innehållsmässigt, men bara en av dem får plats i guldstandarden, vilket gör att de sammanfattningar som plockar ut ”fel” mening kommer få ett lägre betyg (Jing et al. 1998).

Ett annat yttre sätt att utvärdera sammanfattningar på föreslogs av Morris, Kasper & Adams 1992. De ville testa hur bra sammanfattningar fungerade som ersättning för originaltexter i det amerikanska testet Graduate Management Aptitude Test, en motsvarighet till det svenska högskoleprovet. I sitt test lät de olika personer läsa olika sorters sammanfattningar, dels några extraktionsbaserade sammanfattningar som plockade ut meningar på måfå, dels två extraktionsbaserade sammanfattningar som använde en enkel algoritm för att plocka ut 20 respektive 30 procent av de meningar som var mest relevanta för ämnet, andra fick ett abstract skrivet av människor, några läste originaltexterna och en grupp fick ingen text utan fick bara gissa mellan de fem olika svarsalternativ som fanns. Mellan de sammanfattningar som hade en algoritm för att plocka ut de viktigaste

meningarna, abstracts och originaltext uppmättes ingen signifikant skillnad i antal rätt svar. Det fanns dock en skillnad, om än ej statistiskt signifikant, där abstracts presterade bäst, följt av originaltext och sedan de två extraktionssammanfattarna i placering efter längd, där alltså den trettioprocentiga sammanfattningen presterade bättre än den tjugoprocentiga. (Morris et al. 1992)

Metod

Upplägget för denna studie är i form av en fallstudie. Det innebär att studien är utformad med ett enskilt fall för att kunna säga något om andra liknande förekomster i världen. För att testa hur bra extraktionssammanfattningar är utformades således ett test för att undersöka hur mycket information som finns kvar i sammanfattningarna gjorda av CogSum. Läsförståelsetestet från högskoleprovet vårterminen 2011 användes som underlag, det för att de texterna och frågorna till är utformade för att testa hur bra personer är på att förstå det de läser, samt för att diskriminera mellan låg och högpresterande individer. En annan tanke var att personer som läser på universitet skulle vara de som var med i studien, det för att de dels är vana vid att läsa texter för att plocka ut viktig information, samt att högskoleprovet är tänkt att fungera som en bas för hur bra olika individer kommer att klara av högskole- och universitetsstudier. Att testet från 2011 användes berodde på att så få som möjligt på universitet skulle ha gjort testet innan. Data samlades in för två projekt samtidigt, dels för detta projekt men också för (Sandin 2012). Det gjorde att testet blev något längre än det hade behövt vara, eftersom en del till lades till för varje försöksperson, men fördelen blev att mer data från fler försökspersoner kunde samlas in eftersom då två personer kunde göra experimentet oberoende av varandra. För att få två olika baser, samt att alla personer skulle göra lika mycket, användes de fem texter som användes i högskoleprovet vårterminen 2011 (appendix 4). För fyra av texterna producerades en extraktionsbaserad sammafattning (appendix 5), en omskrivning samt behölls originalet till alla texterna. Den femte texten användes som ett invänjningstest, alla fick där läsa originaltexten och svara på tillhörande frågor, det för att alla skulle vara bekanta med testet innan de började.

Försöksdeltagare

De försöksdeltagare som ingick i experimentet var uteslutande studenter, det valdes för att det inte skulle bli för stor skillnad mellan hur mycket personer vanligtvis läser texter för att leta efter information. Det var totalt 60 försöksdeltagare med en medelålder på 22.6 år. Av dessa var 22 stycken kvinnor och 38 stycken män. Det var 13 personer som inte hade gjort högskoleprovet någon gång, och ingen av de som var med i testet hade gjort högskoleprovet VT 2011. 28 av försöksdeltagarna gick samma utbildning, kognitionsvetenskap, resterande gick på olika program på Linköpings universitet. Försöksdeltagarna samlades in med en blandning av bekvämlighetsurval där många av de som deltog i testet på något sätt var bekanta med dem som genomförde testen, och snöbollsurval då vänner och bekanta också tipsade om andra som skulle kunna tänka sig att ställa upp i experimentet.

Utformning av frågeformulär.

Två olika formulär, utöver de frågor som tillhörde varje enskild text, skapades. Dels en enkät som lades före hela testet med olika bakgrundsfrågor om dels egen uppskattad läsförmåga samt bakgrundsfrågor som vilken utbildning och kön, samt om man tidigare gjort högskoleprovet för att kontrollera så att de inte tidigare gjort högskoleprovet från VT 2011 (appendix 1). Efter varje text sammanställdes också ett antal frågor som var tänkta att mäta den subjektiva upplevelsen av texten de precis läst. Dessa frågor var baserade på likertskalor och gick från 1-7 där ett var håller inte med påståendet alls och sju var håller med helt (appendix 2). Frågorna testades i två pilottest för att säkerställa att de inte skulle tolkas på olika sätt, och så att inga av de frågor som ställdes kändes oklara eller svåra att svara på.

Pilottest.

Två pilottest utfördes för att försöka hitta brister i metoden och kunna förändra dessa. I båda dessa test utfördes hela testet av försöksdeltagaren, och båda försöksledarna var på plats för att kunna diskutera proceduren efteråt, det för att få en blick över hur lång tid testet skulle ta samt vilka frågor som var svåra att besvara. Efter varje del i pilottestet ställdes frågor till försöksdeltagaren om hur denne upplevde de olika formulären, om det var något som var oklart eller om det kändes som att något annat borde förändras. I våra olika formulär försökte vi fånga upp olika subjektiva upplevelser av texternas utformning. Dessa lades det stor vikt vid att göra ordentligt, och under det första pilottestet kom det upp ett par synpunkter på de frågor vi ställde. Efter första pilottestet fördes därför en diskussion med handledare och den andra försöksledaren, Sandin 2012, vilket ledde fram till ett par förändringar i formulären. Det kom också fram att försökspersonen under testet ändrade sitt upplägg för att besvara frågorna och läsa texten. På grund av detta valdes ett system ut för att få alla försöksdeltagare att utföra testet i samma ordning, alltså se till att de inte läste frågorna först ena gången och texten först nästa gång. Det andra pilottestet flöt på bra och efter det testet gjordes inga ytterligare förändringar till testets utförande.

Förändringar i frågeformulär ett.

I första upplagan av förenkäten fanns två frågor om hur mycket olika personer läser, dels ”jag läser mycket studielitteratur” och ”jag läser mycket på fritiden”. Dessa frågor ämnade till att fånga upp att, även om man inte läser mycket skönlitteratur kan man läsa mycket studielitteratur eller tvärtom. Dessa frågor tolkades dock som en jämförelse med andra på det program man läser eller mellan de kompisar man har, och gav därför inte något bra resultat. Frågorna slogs ihop till en enda fråga ”Jag läser mycket akademiska texter, skönlitteratur, bloggar, nyheter etc.” för att försöka fånga upp att man kan läsa mycket även om det inte är skönlitteratur eller studielitteratur.

Frågan ”jag har bra läsförmåga” ändrades till ”jag brukar förstå det jag läser” och ”jag läser långsamt”. Det gjordes för att försöka bättre fånga upp det första frågan var ämnad till, men göra den mindre tvetydig, därför delades den upp till två frågor som är lättare att tolka och därmed förhoppningsvis enklare att besvara.

”Jag tycker det är ansträngande att läsa” lades också till för att kunna återkoppla till de frågorna som försöksdeltagarna fick besvara efter varje text. Där fanns redan frågan ”jag tycker att texten var ansträngande att läsa” och här var tanken att även om man tycker att det är ansträngande att läsa i vanliga fall kanske just denna text inte var det. En fråga om högsta poäng på högskoleprovet lades också till för att se om det fanns något samband mellan att tidigare ha ett högt poäng på högskoleprovet och att svara bra på frågorna, men eftersom det var många som ställde upp i testet som inte gjort högskoleprovet, eller av andra anledningar inte ville svara på frågan, användes inte denna data senare.

Förändringar i frågeformulär två.

I det formulär försöksdeltagarna fick svara på efter varje text de läst gjordes också ett par förändringar. I appendix 2 finns formuläret i sin helhet, här följer de förändringar som gjordes i formuläret från hur det såg ut i första pilottestet. Frågan ”jag tycker att texten har en bra längd” togs helt bort, det eftersom det kan vara svårt att besvara frågan, vad skulle texten ha en bra längd för?

Frågorna ”jag tycker att texten var lätt att förstå”, ”jag tycker att texten tog lång tid att läsa” och ”jag tycker att texten var lätt att läsa” lades till för att kunna återkoppla till frågorna från frågeformulär

ett. Dessa frågor lades till för att kunna återkoppla till att de extraktionsbaserade sammanfattningarna kanske upplevs som exempelvis lättare att läsa än vad försöksdeltagarna annars tycker att det är att läsa, eller att originaltexterna ses som svårare att läsa än man generellt upplever att det är.

”Jag upplever texten som informationsfattig” och ”jag tycker att texten var svår att läsa” ändrades till positiva svarsalternativ, det för att det inte var så många frågor och därför ansågs det inte vara en stor risk att försöksdeltagarna skulle svara utan att läsa frågorna. Tanken var då att skalorna skulle vara lika för alla frågor, för att minska risken att försöksdeltagarna svarar 7 på ett påstående när de menade 1, alltså att de inte höll med alls. Det fanns dock fortfarande negativa frågor med då frågorna ”jag tycker att texten var ansträngande att läsa” och ”jag upplever att texten saknar relevant information för att besvara frågorna” ansågs svåra att ändra på till ett positivt alternativ.

Efter dessa förändringar utfördes ett andra pilottest där det återigen frågades om det var några frågor som var oklara eller svåra att besvara. Efter detta pilottest gjordes inga förändringar i frågeformulären eller i upplägget av testet.

Testet

Testet utformades med grund i Morris, Kasper & Adams studie från 1992 där försöksdeltagare fick läsa de texter och svara på de frågor som kom från Graduate Management Aptitude Test, ungefär motsvarande svenska högskoleprovet. I detta test användes högskoleprovet VT 2011 som grund för de texter och frågor försöksdeltagarna blev presenterade. Upplägget på testet var att försöksdeltagarna först fick en kort introduktion för vad som skulle hända, att det var frivilligt och att de närsomhelst kunde avbryta testet. Därefter fick de svara på några allmänna frågor om sig själva, exempelvis ålder och utbildning, samt om de gjort högskoleprovet tidigare för att se så att inte någon gjort högskoleprovet vårterminen 2011 (appendix 1). De fick även uppskatta bland annat sin egen läsvana och hur bra de tyckte att de var på att läsa (appendix 1), flera av dessa frågor kopplades senare ihop med frågor från enkäterna med subjektiva skattningar av hur de olika texterna uppfattades. Därefter fick försöksdeltagarna först läsa de frågor som tillhörde text nummer tre, detta för att de skulle vara presenterade för vad de skulle leta efter i texten. När de läst klart frågorna tog försöksledaren ifrån dem frågorna och gav dem text tre från högskoleprovet. Försöksdeltagarna fick sedan läsa text nummer tre från högskoleprovet VT 2011 i originalform, under tiden de läste texten tog försöksledaren tid på hur lång tid det tog att läsa texten. När de läst klart fick försökspersonerna tillbaka frågorna de läst innan texten för att denna gång besvara dem, här tog försöksledaren varvtid så att även tiden för hur lång tid det tog att svara på frågorna uppmättes (originaltext och frågor till text ett finns i appendix 3, 4). När de hade svarat på frågorna som tillhörde texten stoppade försöksledaren tiden och förde ner den på formuläret med frågor som försöksdeltagaren svarat på. Därefter fick försöksdeltagarna svara på de efterföljande frågor som sammanställts vilka var subjektiva skattningsfrågor om hur de upplevde texten (appendix 2). Alla försöksdeltagare gick igenom denna procedur för text tre, detta för att de skulle få vänja sig vid testets upplägg. Det var av vikt att även de som inte gjort högskoleprovet någon gång skulle få se hur det går till, samt att de som gjort det skulle förstå att det var en liten skillnad i upplägget då de alltid skulle läsa frågorna först och därefter bli presenterade texten. Att text tre valdes som första del var för att den var kortast av de texter som ingick i högskoleprovet VT 2011 och därför skulle ge minst skillnad mot en sammanfattning. När försöksdeltagarna var klara med alla delar på text tre fick de göra ett av fyra set, där det ingick en text i originalform, en sammanfattning, en omskrivning (omskrivningarna är den

del som samlades in för Sandin, 2012 och kommer därför inte få en mer ingående beskrivning i denna rapport), samt en del där man fick gissa på frågorna till den sista texten. Dessa set såg alltid likadana ut, alltså de som läste originalet för första texten läste en omskrivning för andra texten, en sammanfattning för den tredje och gissade på den fjärde. Detta kan ses i tabell 1.

Tabell 1

Försöksperson	originaltext	sammanfattning	omskrivning	gissa
1-15	1	2	3	4
16-30	2	3	4	1
31-45	3	4	1	2
46-60	4	1	2	3

Vilka försökspersoner som läser vilket set av texter.

Att försökspersonerna gissade på sista delen betyder att de inte presenterades någon text, utan enbart fick de tillhörande frågorna från högskoleprovet, och utifrån dessa skulle ringa in det alternativ som de ansåg passade bäst in. Varje försöksledare tog två set var, och hade som mål att samla ihop 15 försöksdeltagare för varje set. Det blev alltså totalt 60 försöksdeltagare i testet. För varje försöksdeltagare kastades ordningen om så att det inte skulle vara så att den text som låg först eller sist alltid fick sämre resultat än de andra. Om originalet av text ett låg först för försöksdeltagare ett låg den således som nummer två för försöksdeltagare två och som nummer tre för försöksdeltagare tre. För varje del i testet togs det tid på hur lång tid det tog att läsa texten, samt hur lång tid det tog att svara på frågorna om texten. Det gjordes för att kunna säga någonting om ifall man sparar tid på att läsa exempelvis en sammanfattning eller om det är så att mycket innehåll förloras så att det ändå tar lika lång tid totalt att försöka leta efter den information man är ute efter för att kunna besvara frågorna. Den del där man enbart gissade på frågorna användes för att skapa en bas för hur mycket man kan tänka sig att personer svarar rätt om de inte har fått någon ytterligare information om texten, det för att se så att inte förkunskaperna inom ett visst ämne skulle spela stor roll för hur bra man svarar på de efterföljande frågorna. Det skulle också kunna vara så att vissa alternativ utesluts på grund av tidigare eller senare frågor och skulle då göra det något lättare att gissa på de andra frågorna. För den del där försöksdeltagarna gissade svarade de inte på de allmänna frågorna om texten då de inte fått någon text att läsa.

Det som mättes upp under testet var alltså antal rätt svar på högskoleprovdelen, hur lång tid det tog att läsa de olika texterna och hur lång tid det tog att svara på de olika delarna samt de subjektiva skattningarna om textens kvalitet från enkäterna efter varje del.

Resultat

För att testa resultatet gjordes uträkningar med hjälp av IBM statistics SPSS 19. Nedan följer de resultat från de objektiva delarna, tid det tog att läsa och antal rätt svar på provet, samt från den subjektiva delen, alltså enkäterna om textens kvalitet. Alla resultat kan antas vara normalfördelade.

Skillnader i antal rätt svar.

För de fyra olika texterna sammanslaget uppmättes skillnader i antal rätt svar med en inomgruppsANOVA. Medelvärdet för de fyra olika texterna tillsammans gav ett medel på 2.62 rätt, eller 65.5% för originaltexterna, 2.2 rätt, eller 55% för sammanfattningarna och för att gissa blev det i snitt 1.3 rätt, eller 32.5 % per text, Huynh-Feldt korrigerat : F(1.86, 109.77)=30.735 p <0.01 η^2= .34. Ytterliggare beskrivning finns i tabell 2.

Tabell 2

Texttyp	Antal rätt	Rätt i procent
Originaltext	2.62	65.5
Sammanfattning	2.2	55
Gissa	1.3	32.5

Antal rätt svar för de olika texttyperna i snitt och i procent.

SIDAK post-hoc test visar var skillnaderna ligger och visar att det är en signifikant skillnad mellan originaltext och sammanfattning i antal rätt svar, p<0.05, där originaltext är bättre än sammanfattningar. Origianltext är också signifikant bättre än att gissa p< 0.01 och sammanfattningen är signifikant bättre än att gissa p<0.01.

Skillnad i tid det tar att läsa.

Med ett t-test mellan tiden det tar att läsa texterna uppmättes också skillnader. Det tar signifikant kortare tid att läsa sammanfattningarna än originaltexterna, *t(59)=17.73 p<0.01,* med en genomsnittlig tid för sammanfattningarna på 153.9 sekunder medan det tar i snitt 337.6 sekunder att läsa originaltexterna, ytterligare beskrivning av resultatet finns i tabell 3. Det ger att det tar ungefär 45% av tiden för att läsa originaltexterna att läsa sammanfattningarna. Det tar alltså inte 30% kortare tid än för originaltexterna vilket var sammanfattningarnas längd.

Tabell 3

Texttyp	Tid att läsa
Originaltext	337.6
Sammanfattning	153.9

Tid det tar att läsa de olika texttyperna.

Eftersom det tog hälften så lång tid att läsa sammanfattningarna men de var 70% kortare än originaltexterna gjordes ett t-test mellan 30% av originaltexttiderna och sammanfattningstiderna i sin helhet. Detta gav t(59)=9, p<0.001, alltså är det en signifikant skillnad mellan hur lång tid det tog att läsa sammanfattningarna och vad som är 30% av lästiden för originaltexterna.

För tiden det tog att svara på de olika frågorna uppmättes ingen signifikant skillnad, där det tog i snitt 193 sekunder att svara på frågorna efter att ha läst originaltexterna medan det tog i snitt 199 sekunder att svara på samma frågor om man läst sammanfattningarna.

Resultat från de subjektiva skattningsfrågorna.

För att testa de allmänna frågorna och om det var några skillnader i hur försöksdeltagarna upplevde de olika testen utfördes en ANOVA för att jämföra mellan de frågor som ställdes innan testet började och de frågor som ställdes efter varje del i testet. För vissa av de frågor som ställdes efter varje text fanns det ingen bakgrundsfråga att jämföra med och därför gjordes ett t-test mellan dessa frågor. Jämförelser finns i tabell fyra, men utan de frågor som ställdes innan testet.

På förtestet uppmanades försöksdeltagarna att skatta på en likertskala mellan 1-7, ”jag brukar förstå det jag läser” och detta ställdes mot deras skattning på en likadan likertskala efter varje text där de fick skatta ”jag tycker att texten var lätt att förstå”. Här uppmättes skillnader $F(2, 118)=24.577$ $p<0.01$ $\eta^2=0.294$, sfärisitet antaget, och ett SIDAK post-hoc test visade att försöksdeltagarna skattade högre på att de brukar förstå vad de läser än vad de tyckte att de förstod i de originaltexter som de presenterades, $p<0.01$. Samma sak gällde mot sammanfattningen, $p<0.001$, och originaltexten skattades också som lättare att förstå än sammanfattningen $p<0.001$.

Samma procedur som ovan gäller följande frågor:
”Jag tycker det är ansträngande att läsa” och ”Jag upplever texten som ansträngande att läsa”. Här blev resultatet $F(2, 118)=4.344$ $p<0.05$ $\eta^2=.069$, sfärisitet antaget, där ett SIDAK post-hoc test visar att skillnaden ligger mellan förtestet, ”alltså jag tycker det är ansträngande att läsa”, och upplevelsen av att sammanfattningen var ansträngande att läsa, $p<0.05$. Ingen signifikant skillnad uppmättes mellan originaltexten och sammanfattningen.

”Jag läser långsamt” och ”jag tycker att det tog lång tid att läsa” gav följande resultat: $F(2.118)=4.346$ $p<0.05$ $\eta^2= .069$, sfärisitet antaget, där ett SIDAK post-hoc test visade att det inte är några skillnader mellan förtestet och originaltexterna. Däremot är det en signifikant skillnad mellan förtestet och sammanfattningarna, $p<0.05$, där sammanfattningarna skattas som att gå snabbare att läsa än vad det går i vanliga fall, samma resultat mättes upp mot originaltexterna där också sammanfattningarna skattas som snabbare än fultexterna, $p < 0.05$.

”Jag har lätt för att läsa” och ”jag tycker att texten var lätt att läsa”, där Mauchley test of sphericity ger $p<0.05$, alltså inte sfäriskt, och Huynh-Feldt korrigering ger $F(1.86, 109.71)=11.876$ $p<0.001$ $\eta^2=.168$. SIDAK post-hoc test visar att skillnaderna ligger mellan förtestet och sammanfattningen, $p<0.001$, samt mellan originaltext och sammanfattning, $p < 0.005$, där sammanfattningen skattas som svårare att läsa båda gångerna.

Följande resultat är från de tvåsidiga t-test som utförts mellan de subjektiva skattningsfrågor om texterna, där det är en jämförelse mellan skattningen på originaltexterna mot sammanfattningarna. Alla är över det kritiska t-värdet och originaltext står först hela tiden, positiv korrelation är alltså att originaltext är högre skattad än sammanfattning och negativ korrelation att sammanfattningarna är högre skattade än originaltexterna. I tabell fyra finns en översikt över resultatet, resultatet från ovanstående ANOVOR är dock inte med i tabellen.

På frågan ”jag tycker att texten ger en bra uppfattning om ämnet” blev resultatet att originaltexterna upplevdes som bättre än sammanfattningarna, $t(59)=7.089$ $p<0.001$. Samma gäller för frågorna ”jag upplever texten som informationsrik”, $t(59)=5.947 p<0.001$, och ”jag tycker att texten har ett bra flyt”, $t(59)=4.977$ $p<0.001$. På frågan ”jag upplever att texten saknar relevant information för att

besvara frågorna” fick originaltext ett lägre skattat värde, vilket innebär att sammanfattningarna upplevdes sakna mer relevant information, $t(59)=-4.847$ $p<0.001$.

Tabell 4

Fråga	Originaltext	Sammanfattning	Signifikansnivå
jag tycker att texten ger en bra uppfattning om ämnet	4.63	3.20	***p < 0.001***
Jag upplever texten som informationsrik	4.70	3.48	***p < 0.001***
Jag tycker att texten har ett bra flyt	4.75	3.63	***p < 0.001***
Jag upplever att texten saknar relevant information för att besvara frågorna	3.25	4.55	***p < 0.001***

Subjektiv fråga, svarsfrekvens för vardera originaltext och sammanfattning samt signifikansnivå från ett tvåsidigt t-test.

Resultatdiskussion

När försöksdeltagarna läste originaltexterna presterade de bättre än med hjälp av de extraktionsbaserade sammanfattningarna när det kommer till att svara på frågorna från högskoleprovet. För originaltexterna var frekvensen av antal rätt svar uppe i 2.62 medan det för sammanfattningarna var i snitt 2.2 rätt per text. Det är en skillnad från Morris, Kasper & Adams studie där det inte var någon signifikant skillnad mellan originaltexterna och sammanfattningarna, även om de visade att en liten skillnad fanns. Att resultatet i denna studie är signifikant kan bero på ett större urval av personer för de olika delarna, då det i denna studie användes 60 personer för varje betingelse medan det i Morris Kasper & Adams studie ingick 24 personer för varje betingelse. Det var även signifikant lättare att svara rätt på sammanfattningarna än vad det var att gissa vilket visar att sammanfattningarna ändå behåller mycket av den relevanta informationen för uppgiften.

Skillnaden mellan sammanfattningarna och originaltexterna i procent var 10.5 procent i antal rätt. Det kan jämföras med att det tog 55 procent mindre tid att läsa sammanfattningarna. När personer måste läsa igenom mycket information på kort tid kan det kanske vara ett byte man är villig att göra. Att det tar 45 procent av den tid det tar att läsa originaltexterna för att läsa sammanfattningarna och inte 30 som var sammanfattningarnas längd kan bero på flera saker. En viktig del är förmodligen att de försökspersoner som var med i studien inte är vana att läsa texter av det slaget, där meningar rycks ur sitt sammanhang för att dra ner på textmassan. Det skulle kunna finnas en inlärningsfaktor om de presenterades för flera liknande texter innan detta test. Det kan också vara så att man stakar sig mer om man inte får meningarna i sitt sammanhang, och därför blir tvungen att läsa samma sak flera gånger. Detta skulle kunna undersökas med exempelvis ögonrörelsekamera, men får lämnas till vidare forskning. Det är dock fortfarande så att tiden det tar att läsa texterna kortas ner med hjälp av extraktionsbaserade sammanfattningar, detta har också visats i andra studier (Mani et al. 1999)

Det tog nästan lika lång tid att svara på frågorna oberoende av vilken text försöksdeltagarna läst. Det skulle kunna betyda att själva beslutsfattandedelen inte blir bättre med hjälp av sammanfattningarna än med originaltexterna. Vad det beror på är svårt att säga, det borde vara lättare att hitta den information man är ute efter när textmassan minskar, men det skulle kunna vara så att mycket tid går åt för att leta efter svar i texten som inte finns. Att undersöka varför det tar lika lång tid att svara på frågorna, och inte bara konstatera att det gör det, skulle kunna vara en intressant fråga att besvara i framtiden.

På de allmänna frågorna till varje text visade det sig att originaltexterna ofta skattades som bättre än sammanfattningarna. De upplevdes som lättare att läsa, att de hade bättre flyt och att de innehöll mer information än sammanfattningarna. Det är inget som förvånar så mycket eftersom sammanfattningarna generellt är lösryckta meningar ur sitt sammanhang medan originaltexterna har all kontextuell information tillgänglig. Det är också så att de personer som var med i studien ofta skattade sin egen förmåga som bättre än vad de upplevde att texterna var. Det skulle kunna bero på att texterna som ingår i högskoleprovet är svårare att läsa än vad de är vana vid eller att de kanske inte rör ämnen som intresserar på samma sätt som annat de är vana att läsa. Det skulle också kunna vara ett tecken på att det är svårt att skatta sin egen förmåga och att det därför ofta skiljer sig mellan förenkäten och de frågor som ställdes efter varje text.

Metoddiskussion

I studien ingick enbart studenter från Linköpings universitet. Det kan finnas anledning att anta att de skulle prestera bättre på högskoleprovet än personer som inte har pluggat vidare. Statistik från högskoleverket visar tydligt på att ju högre utbildning, desto högre poäng på högskoleprovet (Högskoleverket 2001). För att inte få en för stor spridning på de deltagare som var med i testet valdes därför enbart studenter ut för att göra testet.

Det visade sig också att sammanfattningarna inte skattas som mer ansträngande att läsa än originaltexterna, det trots att de är just meningar ryckta ur sitt sammanhang. Försöksdeltagarna skattade också dessa texter som att de gick snabbare att läsa än originaltexterna. Detta är ett resultat som dock ska tolkas relativt till den uppgift försöksdeltagarna gjorde. Sammanfattningarna var för alla set den i särklass kortaste texten, vilket gör att de i denna studie inte kan ställas i relation till en text med samma längd fast med ett bättre flyt. Detta är också något som får lämnas till vidare forskning.

Att studien inte enbart fokuserar på en aspekt av texterna, exempelvis antal rätt svar, utan kollar på tre olika aspekter av texterna, alltså tid det tar att läsa, antal rätt svar samt subjektiva skattningar av texternas estetiska del gör att validiteten för studien ökar. Det faktum att alla deltagare gjorde alla delar i samma ordning gör också att validiteten ökar, då det annars hade kunnat bli så att de som läser frågorna först svarar bättre, eller att de som läser texten först svarar bättre. Detta gör att en störvariabel kontrolleras för vilket styrker validiteten, men det kan också ha gjort att personer som föredrar att göra på ett annat sätt svarar sämre än de skulle gjort annars. Detta ansågs dock inte möjligt att genomföra i denna studie men skulle kunna vara något att studera vidare i framtiden.

Vikten av högskoleprovets utformning är inte heller att förringa i resultatet. Läs-delen på högskoleprovet är utformat för att fånga upp läsförståelse, och frågorna är inte utformade för att visa att personer kan ta ut det viktigaste ur en text, utan många frågor är svåra att besvara även om man har all information från texterna tillgänglig. Det kan man se i resultatet där deltagarna läst originaltexterna i snitt endast svarar 2.62 rätt i snitt av 4 möjliga. De flesta extraktionssammanfattare arbetar med att försöka plocka ut den viktigaste informationen ifrån de texter den sammanfattar(Chatterjee & Mohan 2007, Mani et al. 1999), detta gör även CogSum (Smith & Jönsson 2011 A), därför är högskoleprovet kanske inte helt rättvist för att utvärdera det.

Studien visade även att det är något lättare än 25 procents chans att svara rätt på de frågor som är på högskoleprovet även om man enbart presenteras för frågan och svarsalternativen, 32.5 procent. Med det i åtanke hade det kunnat vara en fördel att ha med en följdfråga till varje fråga från högskoleprovet i stil med ”Jag upplever att svaret fanns i texten”, för att kunna få en bild av hur ofta försöksdeltagarna ansåg att de gissade på en fråga från sammanfattningarna och originaltexterna. Det hade kunnat ge upphov till andra intressanta resultat för studien, kanske upplever försöksdeltagarna att de gissar lika ofta för originaltexterna som för sammanfattningarna, eller kanske upplever försöksdeltagarna att de gissar mer med sammanfattningarna, men gör det på en bättre grund än om de inte fått någon text alls. Det får dock lämnas till vidare forskning.

Ett problem som upptäcktes med sammanfattningarna från CogSum är att punkter används som meningsslut. För till exempel lagtexten, som hade många förkortningar i stil med ”i kap. 3”, skalas resten av meningen efter ”i kap.” bort, vilket gav ett par meningar som inte gav så mycket information som hade kunnat önskas. Detta hade kanske kunnat lösas med någon form av lexikon för

vad som är vanliga förkortningar i svenskan, så att inte meningar plockas ut på fel grunder av sammanfattaren.

Slutsats

I den här studien har det visats att extraktionsbaserade sammanfattningar inte är en fullgod ersättning till originaltexterna från det svenska högskoleprovet. Det visade sig dock att sammanfattningarna är signifikant mycket bättre än att gissa på frågorna, och var endast tio procent sämre än originaltexterna, vilket visar på att mycket information bevaras även när textmassa reduceras med 70 procent. De extraktionsbaserade sammanfattningarna presterar alltså bra i jämförelse med originaltexterna, och det skulle kunna tänkas att om försökspersonerna tidigare fått träna på att läsa några extraktionsbaserade sammanfattningar skulle presterat bättre i testet.

Det visade sig också att det tar lite mindre än hälften så lång tid att läsa sammanfattningarna i jämförelse med originaltexterna, vilket kan jämföras med den tioprocentiga försämringen i antal rätt svar. I situationer där personer ska läsa mycket text under kort tid kan därför dessa sammanfattningar vara ett bra alternativ till originaltexterna. Sammanfattningarna borde också testas i ett liknande test, men ett test som skiljer sig från högskoleprovet så att det handlar om att få ut den viktigaste informationen, vilket inte alla frågor i högskoleprovet är ämnade att fånga upp i läsförståelsedelen. Detta skulle kunna ge upphov till andra intressanta resultat, som kanske visar på att sammanfattningarna bättre bevarar den viktiga informationen, men saknar mycket av det som går att tolka in i en text där all information finns kvar.

Försöksdeltagarna skattade ofta sammanfattningarna som sämre än originaltexterna när det gällde den subjektiva upplevelsen, vilket visar att det fortfarande finns mycket att göra när det kommer till den biten av extraktionsbaserade sammanfattningar. Men samtidigt var försöksdeltagarna inte medvetna om att det var en text skapad på detta sätt och har förmodligen ingen, eller väldigt lite, vana av att läsa texter skapade på detta sätt.

Referenser

Chatterjee, N. Mohan, S. (2007) Extraction-based Single-Document Summarization Using Random Indexing. *19th IEEE International Conference on Tools with Artificial Intelligence*

Chin-Yew, L. Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. *Proceedings of HL+T-NAACL 2003,* 71-78

Hassel, M. (2004). Evaluation of Automatic Text Summarization. *Licentiate Thesis*, 3-20

Hassel, M. (2007). Resource Lean and Portable Automatic Text Summarization. *Doktorsavhandling KTH School of Computer Science and Communication*

Jing, H. McKeown, K. Barzilay, R. Elhadad, M. (1998). Summarization Evaluation Methods: Experiments and Analysis. *AAAI Technical Report SS-98-06*

Mani, I. Firmin, T. Sundheim, B. (1999). The TIPSTER SUMMAC Text Summarization Evaluation. *Proceedings in EACL*

Carlsson, B. (2009). Guldstandarder – dess skapande och utvärdering. *Kandidatuppsats Linköpings Universitet.*

Morris, A. Kasper, G. Adams, D. (1992). The effects and limitations of automated text condensing on reading comprehension performance. *Advances in automatic text summarization* (305-323)

Sahlgren, M. (2005) An Introduction to Random Indexing . *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*

Sandin, J. (2012). Kandidatuppsats. Ej publicerad

Smith, C., Jönsson, A. (2011 A). Automatic Summarization As Means Of Simplifying Texts, An Evaluation For Swedish. *Proceedings of the 18th Nordic Conference of Computational Linguistics (NoDaLiDa-2010)*, Riga, Latvia 2011.

Smith, C. Jönsson, A. (2011 B). Enhancing extraction based summarization with outside word space. *Proceedings of the 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, 2011.

Högskoleverket. (2001). http://www.edusci.umu.se/digitalAssets/57/57755_pmnr172sec.pdf (länk från 2012-06-07)

Appendix

Nedan följer de olika formulär som använts i testet, samt en av de sammanfattningar och originaltexter som använts i experimentet.

Appendix 1

Frågeformulär 1.

Man ____ Kvinna ____

Ålder _____

Utbildning/yrke _____

Vilken termin _____

Jag läser mycket akademiska texter, skönlitteratur, bloggar, nyheter etc.(där 1 är ”håller inte med alls” och 7 är ”håller helt med”)

1 2 3 4 5 6 7

Jag brukar förstå det jag läser. (där 1 är ”håller inte med alls” och 7 är ”håller helt med”)

1 2 3 4 5 6 7

Jag läser långsamt. (där 1 är ”håller inte med alls” och 7 är ”håller helt med”)

1 2 3 4 5 6 7

Jag tycker att det är roligt att läsa. (där 1 är ”håller inte med alls” och 7 är ”håller helt med”)

1 2 3 4 5 6 7

Jag har lätt för att läsa. (där 1 är ”håller inte med alls” och 7 är ”håller helt med”)

1 2 3 4 5 6 7

Jag tycker det är ansträngande att läsa. (där 1 är ”håller inte med alls” och 7 är ”håller helt med”)

1 2 3 4 5 6 7

Har du tidigare gjort högskoleprovet? _____

Om ja.

När? _____

Hur många gånger? _____

Högsta poäng på högskoleprovet? _____

Appendix 2

Frågeformulär två.

Jag tycker att texten var lätt att förstå? (där 1 är ”håller inte med alls” och 7 är ”håller helt med”)

1 2 3 4 5 6 7

Jag tycker att det tog lång tid att läsa? (där 1 är ”håller inte med alls” och 7 är ”håller helt med”)

1 2 3 4 5 6 7

Jag tycker att texten ger en bra uppfattning om ämnet. (där 1 är ”håller inte med alls” och 7 är ”håller helt med”)

1 2 3 4 5 6 7

Jag upplever texten som informationsrik. (där 1 är ”håller inte med alls” och 7 är ”håller helt med”)

1 2 3 4 5 6 7

Jag tycker att texten är lätt att läsa. (där 1 är ”håller inte med alls” och 7 är ”håller helt med”)

1 2 3 4 5 6 7

Jag tycker att texten har bra flyt. (där 1 är ”håller inte med alls” och 7 är ”håller helt med”)

1 2 3 4 5 6 7

Jag tycker att texten är ansträngande att läsa. (där 1 är ”håller inte med alls” och 7 är ”håller helt med”)

1 2 3 4 5 6 7

Jag uppfattar att texten saknar relevant information för att besvara frågorna. (där 1 är ”håller inte med alls” och 7 är ”håller helt med”)

1 2 3 4 5 6 7

Övriga kommentarer?

Appendix 3

Frågor från högskoleprovet

1. Vad anser Arne Jarrick om Laqueurs skildring av den viktorianska tiden, enligt texten?

A Att den är alltför akademisk och svårtillgänglig.

B Att den är alltför förenklad och ensidig.

C Att den är alltför naturvetenskapligt inriktad.

D Att den är alltför högtravande och positiv.

2. Jarrick har en viss uppfattning om kärleken som han vill bevisa. Hur ser recensenten på de bevis som Jarrick hänvisar till?

A Bevismaterialet motsäger snarare Jarricks uppfattning.

B Bevismaterialet utgörs i för hög grad av andra litterära källor.

C Bevismaterialet är begränsat och egendomligt utvalt.

D Bevismaterialet är allmänt och saknar historisk förankring.

3. Arne Jarrick kritiserar enligt texten Thomas Laqueur för bristande konsekvens. Vilken av Laqueurs teser gäller denna kritik?

A Tesen att kulturen i samhället i huvudsak är en biologisk produkt.

B Tesen att vetenskapens rön speglar den rådande kulturen.

C Tesen att jämlikhetstanken föddes under 1700-talet.

D Tesen att moralen och kulturen tenderar att förstärka varandra.

4. Om vi utgår från recensentens beskrivning av innehållet i Jarricks bok, vilken alternativ titel skulle då passa boken bäst?

A ”Kärleken med stort K”.

B ”Kärleken till kvinnan”.

C ”Kärleken i skönlitteraturen”.

D ”Kärleken till kulturen”.

Appendix 4

Kärlekens makt och tårar (original)

Arne Jarrick har skrivit Kärlekens makt och tårar, en bok på 300 sidor. Där behandlas kärleken i rättegångshandlingar från 1700-talet. Vidare diskuteras fyra olika faser i synen på kvinnan; ett stort kapitel innehåller analyser av en rad skillingtryck.

Jarrick vill bevisa att kärleken, eller "det romantiska känslospråket", fanns redan före romantiken. Han vänder sig mot "konstruktivisterna" och särskilt mot historikern John Gillis, som skrivit att kärleken saknar "tidlös essens" och att den är en "kulturell konstruktion".

Jarricks extremt kortfattade referat av Gillis är svårt att begripa. De flesta av oss tror nog spontant att kärleken saknar tidlös essens. Vad är det för konstigt med det? Hävdar man motsatsen bygger det väl på att man ger den vackraste visan om kärleken en religiös dimension. Men inte ens du och jag som älskar så hett och unikt är ju odödliga, om man ser strikt vetenskapligt på det hela. Och har vi ett ögonblick inbillat oss att lågan mellan oss, så starkt lysande, är tidlös i någon mening? Då vore den ju inte vår. Och då ägde den knappast den sårbarhet som gör den så sensationell.

Men därav följer knappast att vår kärlek är en kulturell konstruktion. Under inga omständigheter vår kärlek! Och inte heller skorstensfejare Johan Arwid Horns kärlek till Christina Funck, som Arne Jarrick letat fram från konsistorieprotokollet från 1735, lika litet som en hel rad andra vittnesmål från samma tid om autentiska känslor, som han också tycker har stort bevisvärde.

Det finns naturligtvis från historisk tid hur många vittnesmål som helst om känslor som vi lätt som en plätt identifierar som kärlek av något slag och som till sin kärna knappast kan vara enbart kulturbetingade. Ändå kan det vara rimligt att tala om kärleken som en kulturkonstruktion, nämligen om detta allmänbegrepp behandlas som en egen och oberoende existens och blir ett slags personifikation som hävdar sig på egen hand bortom alla de enskilda kärlekarna. En "kärlek", alltså, som gör anspråk på att vara likadan i alla tider och kulturer, eller som under namn av Eros tillmäter sig vad Arne Jarrick gärna ser hos den: tidlöshet.

Kärleken är långt äldre än "den höga romantiken", påpekar Jarrick. Den har långa anor, stryker han under, och tar därmed till en bild som förvandlar kärleken från ett samlande namn på likartade sensationer genom tidernas lopp till en essens som flyter genom historien och med sin eldfängdhet tänder exakt likadana brasor här och där, kort sagt: en kulturell konstruktion, som nog kan hävda sig ganska långt men också svika oss grymt. Det händer ju att man baserar sin känsla på en tillit till vad man hört om Kärleken med stort K och inte på den faktiska interaktionen med den man eller kvinna som livet fört i ens väg. Sådant kan ge obehagliga överraskningar!

Det som gör det svårt att veta riktigt hur man ska uppfatta Jarrick, det är att hans motståndare – konstruktivisterna – framställs som så enögda att man inte förstår att de är värda allt krutet.

När han bevisar att det fanns romantiska känslor före romantiken – varmed uppenbarligen menas högromantiken vid 1800-talets början – så är det bara alltför självklart. Tänk bara på Höga visan, på Sapfos och Propertius dikter, på Abélard och Héloïse, Petrarca och Laura, Shakespeares sonetter och tusen vittnesmål från alla tider om förälskelser av olika slag, kärlekssjuka, svartsjuka, crime

passionnel! Man frågar sig förundrad varför författaren inte använder sig av detta överväldigande material i stället för att leta upp några konsistorieprotokoll från 1700-talets mitt.

Inte kan det väl vara för att dessa – till skillnad från exempelvis berättelsen om Abélard och Héloïse – skulle vara av språk och kultur totalt obesmittade och därmed de enda dyrbara bevisen på romantiska känslor före romantiken? Den givna invändningen är då att man knappast behöver vara konstruktivist för att hävda att sådana "absoluta", av språk och kultur oberoende, känslor knappast kan urskiljas. Även kroppsspråket är ett språk, av kulturen berört.

Kanske är det så att de mest äkta uttrycken för känslor är de språkligt mest avancerade. Språket både förfinar och hetsar kärleken. Och tvärtom. Det är också med och skapar "äkta" kärlek, bland annat eftersom det är en oavskiljbar del av interaktionen mellan de älskande.

Arne Jarricks nästa off er är Thomas Laqueur, som 1990 gav ut Making Sex (svensk titel: Könens uppkomst), vari skildras hur den anatomiska vetenskapen så småningom lämnar enkönsmodellen för tvåkönsmodellen och hur man ska tolka detta. Kvinnan hade samma slags könsorgan som mannen enligt Aristoteles, vagina var en inverterad, inåtvänd penis. Hennes "vätskor var otillräckligt upphettade och därför var hon ett mindre fullkomligt människoväsen än han".

Thomas Laqueur försöker enligt Arne Jarrick påvisa det kulturellt villkorliga i alla biologiska distinktioner. Under renässansen såg man mannens och kvinnans könsorgan i enlighet med Aristoteles, ty Aristoteles var en auktoritet för all vetenskap. Sedan – på 1700-talet – när jämlikhetstankar började slå igenom såg man i enlighet med det då kulturellt gillade sättet, jämlikhetsidealen, att kvinnan hade helt egna organ, hon var inte ett slags sämre man. Under 1800-talets första hälft drevs enligt Laqueur tvåkönsmodellen så långt att man inom medicinen hävdade att det rådde en fundamental skillnad mellan mannens och kvinnans njutning. Kvinnan hade inget nöje av samlaget. Detta vetenskapliga resultat var också kulturellt betingat: avsexualiseringen av kvinnan passade utmärkt in på 1800-talets viktorianska moral.

Detta är schematiskt och grovt Laqueurs ståndpunkt. Jarrick går hårt åt den. Dels är hans antagonist inte konsekvent, han erkänner att vetenskapliga resultat ibland är primära. Dels kan man påvisa att den viktorianska moralen inte alls uteslöt helt andra "vetenskapliga resultat" och en helt annan syn på kvinnan. 1800-talet var, betonar Jarrick, inte alls så homogent som det kan verka hos Laqueur.

Appendix 5

Kärlekens makt och tårar (Sammanfattning)

Han vänder sig mot "konstruktivisterna" och särskilt mot historikern John Gillis, som skrivit att kärleken saknar "tidlös essens" och att den är en "kulturell konstruktion".

Vad är det för konstigt med det? Hävdar man motsatsen bygger det väl på att man ger den vackraste visan om kärleken en religiös dimension. Men inte ens du och jag som älskar så hett och unikt är ju odödliga, om man ser strikt vetenskapligt på det hela. Då vore den ju inte vår. Och då ägde den knappast den sårbarhet som gör den så sensationell.

Ändå kan det vara rimligt att tala om kärleken som en kulturkonstruktion, nämligen om detta allmänbegrepp behandlas som en egen och oberoende existens och blir ett slags personifikation som hävdar sig på egen hand bortom alla de enskilda kärlekarna. En "kärlek", alltså, som gör anspråk på att vara likadan i alla tider och kulturer, eller som under namn av Eros tillmäter sig vad Arne Jarrick gärna ser hos den: tidlöshet.

Det händer ju att man baserar sin känsla på en tillit till vad man hört om Kärleken med stort K och inte på den faktiska interaktionen med den man eller kvinna som livet fört i ens väg.

Det som gör det svårt att veta riktigt hur man ska uppfatta Jarrick, det är att hans motståndare ? konstruktivisterna ? framställs som så enögda att man inte förstår att de är värda allt krutet.

Inte kan det väl vara för att dessa ? till skillnad från exempelvis berättelsen om Abélard och Héloïse ? skulle vara av språk och kultur totalt obesmittade och därmed de enda dyrbara bevisen på romantiska känslor före romantiken?

Och tvärtom. Det är också med och skapar "äkta" kärlek, bland annat eftersom det är en oavskiljbar del av interaktionen mellan de älskande.

Under 1800-talets första hälft drevs enligt Laqueur tvåkönsmodellen så långt att man inom medicinen hävdade att det rådde en fundamental skillnad mellan mannens och kvinnans njutning.

Dels kan man påvisa att den viktorianska moralen inte alls uteslöt helt andra "vetenskapliga resultat" och en helt annan syn på kvinnan.