Att upptäcka plagiering

FELICIA DENBU WILHELMSSON
och EMILIA HILLERT

KTH Datavetenskap
och kommunikation

Examensarbete
Stockholm, Sverige 2011

Att upptäcka plagiering

F E L I C I A D E N B U W I L H E L M S S O N
o c h E M I L I A H I L L E R T

Examensarbete i datalogi om 15 högskolepoäng
vid Programmet för datateknik
Kungliga Tekniska Högskolan år 2011
Handledare på CSC var Johan Boye
Examinator var Mads Dam

URL: www.csc.kth.se/utbildning/kandidatexjobb/datateknik/2011/denbu_wilhelmsson_felicia_OCH_hillert_emilia_K11043.pdf

Kungliga tekniska högskolan
Skolan för datavetenskap och kommunikation

KTH CSC
100 44 Stockholm

URL: www.kth.se/csc

"Art is either plagiarism or revolution"–Paul Gauguin

Abstract

The detection of plagiarism

This essay deals with the topic of plagiarism and is based on the question “is it possible to plagiarize someone else‘s work and then get away with it?”. The essay describes the context of plagiarism as well as the popular plagiarism checking tools. This is followed by a comparison and analysis of some selected tools. From the results it was found, that it in many cases is possible to “fool” some plagiarism checking tools. Finally, based on the results obtained, a draft of a tool that modifies a text in order to avoid being “caught” in a plagiarism checking tool is presented.

Referat

Denna uppsats behandlar ämnet plagiering och utgår från frågeställningen “är det möjligt att plagiera någon annans skrivna verk och sedan komma undan med det?”. Uppsatsen går igenom bakgrund om plagiering samt de populäraste plagieringskontrollsverktygen. Därefter följer en jämförelse och analys av vissa utvalda sådana verktyg. Utifrån resultaten konstaterades att det med hjälp av olika metoder i många fall är möjligt att “lura” vissa plagieringskontrollsverktyg. Slutligen ges, utifrån de erhållna resultaten ett förslag på ett verktyg som modifierar texter för att undvika att “fastna” i ett plagieringskontrollsverktyg.

Förord

Då denna uppsats framförallt bygger på undersökningar, har vi valt att ej dela upp de olika delarna oss sinsemellan. Detta innebär att båda är lika delaktiga i samtliga delar.

Vi vill tacka vår handledare Johan Boye för den goda assistansen han givit oss genom hela arbetsgången.

Innehåll

Kapitel 1

Inledning

Denna uppsats behandlar plagiering och utgör rapporten i ett examensarbete inom datalogi, med inriktning på språkteknologi.

1.1 Bakgrund

Internet växer i mycket snabb takt och det är numera möjligt att finna en dator i nästintill varje hushåll. Detta har lett till att mer information finns lättillgänglig för allmänheten, vilket även öppnar upp möjligheten att kunna plagiera någon annans verk.

I ett försök att motverka plagiering, har det under de senaste åren uppkommit en mängd hemsidor och program som erbjuder plagiatkontroll. Dessa verktyg är främst till för skolor, där problemet är som mest utbrett. Verktygen fungerar något olika, men gemensamt är att de alla har tillgång till fakta och information som finns på Internet, i böcker och artiklar, men även i ett stort antal inlämnade studentuppsatser. Dessa jämförs sedan med de nya uppsatser som studenter lämnar in, och på detta vis kan man kontrollera att uppsatserna ej är plagierade.

1.2 Syfte

Syftet med denna uppsats är att noga undersöka plagieringskontrollsverktyg, samt finna eventuella svagheter hos dessa för att "lura" dem. Viktigt att poängtera är att ett verktyg ej kommer att implementeras i detta projekt.

1.3 Disposition

Uppsatsen är uppdelad i fyra delar, som alla beskrivs kortfattat nedan.

Del I – Bakgrundsteori Denna del ger läsaren all den information och kunskap gällande plagiering, språkteknologi, grammatik samt plagieringskontroller som krävs för att vidare förstå den utförda undersökningen.

Del II – Undersökning med resultat och analys Denna del beskriver grundligt tillvägagångssättet i samtliga tester som utförts, där texter har modifierats manuellt i ett försök att"lura" plagieringskontrollsverktygen. Efter varje test följer resultat av testet samt analys av resultaten.

Del III – Att "lura" Turnitin I denna del utnyttjas all insikt och kunskap som erhållits från de utförda testerna, och ett förslag på implementation av ett verktyg för att "lura" plagieringskontrollsverktyg presenteras.

Del IV – Analys Här diskuteras uppsatsens resultat, och förslag på förbättringar av ett plagieringskontrollsverktyg ges.

Del I

Bakgrundsteori

Kapitel 2

Plagiering

2.1 Definitionen av plagiering

Ordet plagiat kommer från det latinska "plagarius" som betyder kidnappare, förövare, plundrare, litterär tjuv. Plagiat är inte ett rättsligt begrepp, det vill säga att det finns ingen definition av begreppet i svensk lagstiftning [17]. Däremot har skolor lokalt definierat regler angående åtgärder vid upptäckt av plagiering. Eleverna inom varje skola kan läsa om reglerna i skolans egen hederskodex. Finns det misstankar eller klara bevis till att eleven har plagierat överlämnas denne till skolans disciplinnämnd och kan få en varning eller avstängning.

Det finns inget rättsligt begrepp för plagiering, dock ger Nationalencyklopedin följande definition av begreppet plagiering:

"Konstnärlig imitation eller stöld, i fråga om litteratur ofta direkt avskrivning, gjord av någon som ger sig ut för att vara upphovsmannen. Plagiat av till exempel litterära och konstnärliga verk samt skyddad formgivning är i regel juridiskt otillåten." [6]

Plagiering förekommer i flera olika former:

- Att lämna in någon annans arbete i sitt namn.
- Att inte ange referens då:
  - Meningar eller stycken är direkt tagna från någon annans arbete.
  - Någon annans arbete är omformulerat med egna ord.
  - Någons annans idéer, metoder eller data används.

- Att den egna texten är för lik originaltexten, trots korrekta referenser till originalarbetet:
  - Då endast ett fåtal ord byts ut mot synonymer.
  - Att använda sig av originalarbetets formuleringar utan att markera med citattecken.
  - Då det är frågan om en direkt översättning från ett språk till ett annat och detta inte framgår [13].
- Att ge felaktig eller ej tillräcklig information om varifrån materialet kommer. [12]

2.2 Problemen med plagiering

Plagiering är framförallt ett problem inom skolan, där det har blivit allt vanligare att studenter ser på plagiering som en enkel genväg, alternativt som en sista utväg. Detta kan vara en följd av att det blivit lättare att söka efter och få tag i specifik information.

En anledning till varför studenter plagierar kan vara att de aldrig har fått lära sig hur de till fullo ska utnyttja källor på ett korrekt sätt. Det verkar dessutom som att studenter ej anser att det är fel att fuska om det endast rör sig om smärre fusk [7]. Däremot verkar det som att studenter anser att det är en stor skillnad mellan att fuska på en tenta eller fuska på andra sätt, som ofta ses som mindre allvarliga [8]. Vissa studenter plagierar för att orka med alla krav och klara pressen som studier ibland innebär. En annan faktor är den att känna pressen att prestera, att få bra betyg. Vissa engagerar sig också i alldeles för många andra aktiviteter efter skolan, vilket gör att de får tidsbrist och då ser på plagiering som en sista desperat utväg [7].

Kapitel 3

Språkteknologi

Språkteknologi kan och bör användas för att upptäcka plagiering. Dessutom lämpar det sig att använda språkteknologi för att skapa ett verktyg som kan "lura" en plagieringskontroll. Därför följer nedan en kortare introduktion till ämnet språkteknologi.

3.1 Vad är språkteknologi?

Språkteknologi är ett forskningsområde som täcker ett mycket brett spektrum av aktiviteter, och studier inom detta ämne kräver kunskaper inom både dataveten-skap, lingvistik, psykologi och teknik. Det slutgiltiga målet är att möjliggöra för människor att kommunicera med maskiner som har naturliga kommunikationsfärdigheter. För att kommunikationen mellan maskin och människa skall kännas så naturlig och behaglig som möjligt, krävs en djup förståelse dels för den akustiska och symboliska strukturen av språk, dels för de mekanismer och strategier som människor använder för att kommunicera med varandra [2].

Till dessa maskiner räknas system som exempelvis kan tyda, skapa eller känna igen mänskligt språk. Själva tekniken utgörs ofta av ett program eller ett verktyg som använder sig av så kallade språkresurser såsom ordböcker och lexikon. Exempel på sådana tekniker är den rättstavningsfunktion som finns i de flesta ordbehandlingsprogram och översättningsverktyg såsom "Google Översätt".

Språkteknikutvecklingsprocessen går ut på att med tal eller text föra in material i datorn, för att sedan låta datorn tolka och förstå detta material. Datorn får sedan utnyttja den uppnådda förståelsen till att exempelvis översätta eller rättstava en text, eller till att göra en överföring från tal till text.

3.2 Tillämpningsområden

Det finns ett antal olika tillämpningsområden inom språkteknologi, men denna uppsats kommer främst att handla om textanalys, informationshantering samt datasamlingar och lexikon [14]. För den intresserade läsaren finns djupgående information om språkteknologi i (Jurafsky & Martin, 2009)[5].

Kapitel 4

Grammatik

I denna uppsats kommer delar av den svenska grammatiken att användas, dels för att på olika sätt undersöka plagieringskontrollsverktyg, men grammatiska regler kommer även att utnyttjas i ett försök att "lura" dessa verktyg. För att läsaren lättare skall förstå de tester som utförs följer nedan förklaringar av de delar av den svenska grammatiken som det krävs att läsaren behärskar.

4.1 Ordklasser

En ordklass är en gruppering av ord, där orden liknar varandra på det vis att de fungerar på samma sätt. Ordklasser gör det lättare att skilja ord åt och veta hur de ska användas. De fyra ordklasserna som huvudsakligen kommer att undersökas i denna rapport är substantiv, verb, adjektiv och adverb.

4.1.1 Slutna och öppna ordklasser

Ordklasser delas ofta upp i två olika kategorier: slutna och öppna ordklasser. Slutna ordklasser är klasser dit det sällan tillkommer några nya ord. Ett exempel på en sluten ordklass är prepositioner, då det är sällan, för att inte säga aldrig som det skapas nya prepositioner. Några andra slutna ordklasser är pronomen, räkneord, determinerare, konjunktioner och subjunktioner. Substantiv och verb däremot, är exempel på öppna ordklasser, eftersom det dels uppkommer nya sådana men även tillkommer låneord från andra språk. Andra öppna ordklasser är adjektiv, egennamn och adverb. De ordklasser som undersöks i denna rapport är alltså öppna ordklasser. [5]

4.2 Aktiv och passiv form

Verb kan antingen vara aktiva eller passiva. Ett aktivt verb talar om vad subjektet gör, jämfört med ett passivt verb som talar om vad som sker med subjektet. Detta

kan användas för att skriva en mening i antingen aktiv eller passiv form. Med aktiv form menas “X gör Y” och med passiv form menas “Y görs av X”.

4.3 Homonymer och homografer

Det finns en hel del ord som stavas likadant, men har olika betydelser. Homonymer är ord som stavas likadant och även uttalas på samma sätt. Homografer är ord som stavas likadant, men uttalas på olika sätt [15]. Ett ord som är en homonym eller homograf behöver alltså inte tillhöra endast en ordklass.

Kapitel 5

Plagieringskontrollsverktyg

Det finns en mängd olika plagieringskontrollsverktyg på marknaden att tillgå. De flesta av dessa är främst utvecklade för att minska plagiering inom skolan, men används även till annat. Nedan kommer en beskrivning av några utvalda sådana verktyg.

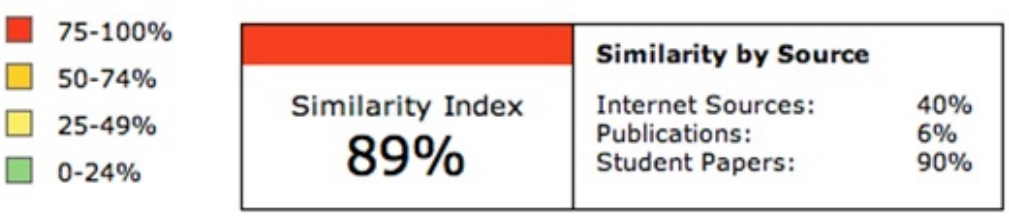
5.1 Turnitin

Turnitin (turnitin.com) är det populäraste och mest använda verktyget på Internet för att kontrollera uppsatsers originalitet och används för tillfället i 126 länder världen över. Det är ett mycket effektivt och pålitligt verktyg och varje dag skickas 150 000 uppsatser in dit. Turnitin accepterar flera olika filformat, dessa är: MS Word (.doc), WordPerfect (.wpd), PostScript (.eps), Portable Document Format (.pdf), HTML (.htm), Rich Text (.rtf) och Plain Text (.txt). Syftet med Turnitin är att förebygga plagiering och lärare såväl som studenter kan skicka in studentuppsatser dit för att granska innehållet.

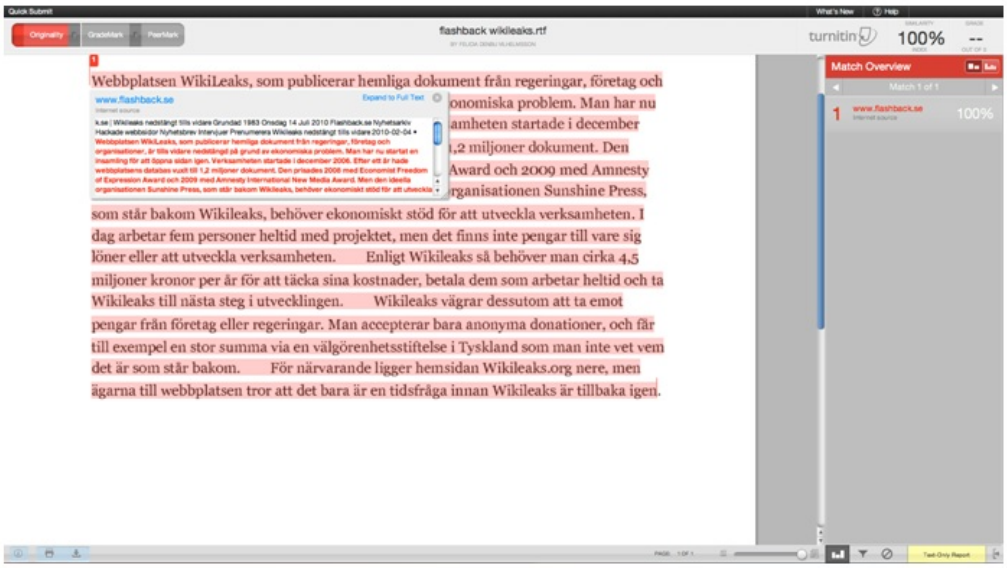
5.1.1 Hur fungerar Turnitin?

Verktyget returnerar en procentsats som anger hur stor del av uppsatsens innehåll som matchar något innehåll i Turnitins databas, se figur 5.1. I figuren visas även hur en text får en viss färg beroende på hur stor matchningsprocent den fick.

Turnitin erbjuder även möjligheten att få se vilken eller vilka delar av uppsatsen som matchade någon annan text i databasen, genom att dessa delar i texten färgmarkeras, se figur 5.2. Där finns det även möjlighet att se en jämförelse mellan den inskickade texten och den text som den matchade, för att se vilka delar som stämmer överens.



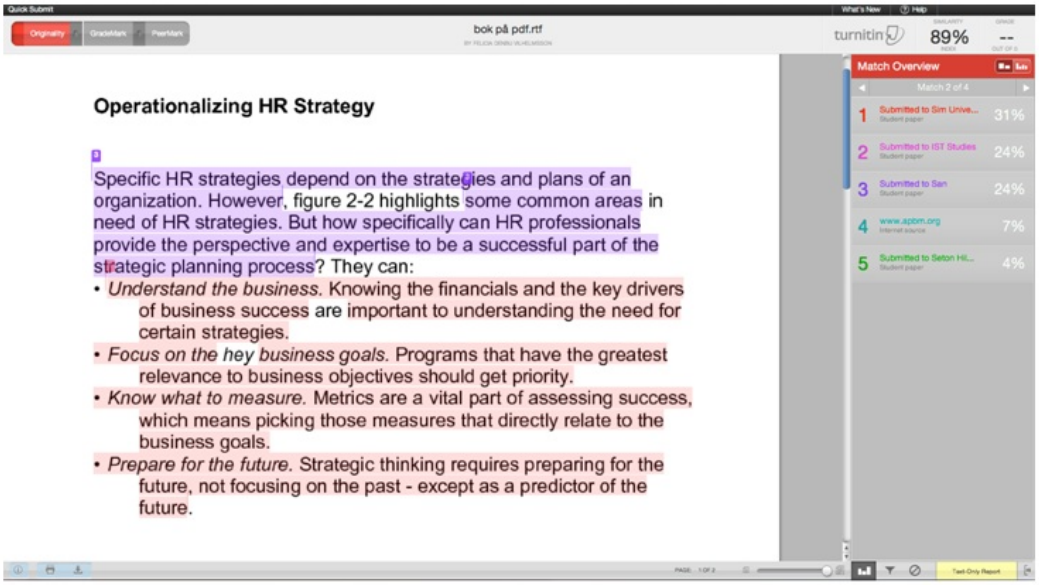
Figur 5.1. De olika gränserna i matchningsprocent hos Turnitin, samt en illustration av hur det ser ut när en text har blivit kontrollerad.



Figur 5.2. En text inskickad till Turnitin där hela texten matchas till en annan källa.

Om en inskickad text matchar flera olika andra källor, så syns detta, då varje källa markeras med en separat färg, se figur 5.3. Text som inte markeras med någon färg har alltså inte matchat någon text i databasen.

Det är viktigt att inse att Turnitin endast kan indikera att en del av en text är mycket lik eller identisk med en annan text. Verktyget kan alltså inte avgöra innebörden av texterna, och därmed är det inte alltid relevanta matchningar som görs. Även om Turnitin jämför texter och hittar matchningar så är det även viktigt att poängtera att detta verktyg aldrig avgör om en student har plagierat eller inte, det är upp till läraren att själv göra.



Figur 5.3. En text inskickad till Turnitin där texten matchas till flera olika källor.

5.1.2 Matchningstekniken

I de databaser som Turnitin använder sig av för att kontrollera uppsatsers originalitet finns miljoner böcker och publikationer, över 14 miljarder websidor och 150 miljoner studentuppsatser. Webbsidorna får de tillgång till via sin egen sökrobot (WebCrawler) som skannar Internet för att kunna indexera innehållet till ett format som går att söka i. Bland denna information finns gammalt arkiverat Internetinnehåll som inte längre finns tillgängligt på Internet. Turnitin har även ett samarbete med de ledande bokförlagen, vilket gör att de har tillgång till biblioteksdatabaser, vanliga böcker och e-böcker, prenumerationsbaserade publikationer och information från sidor på Internet som erbjuder läxhjälp [16].

Turnitin klarar alltså av att matcha exakta plagieringar, men enligt dem själva klarar deras plagieringskontroll av betydligt mer komplicerade saker. De påstår bland annat att de klarar av att identifiera när en mängd av orden i en text har blivit utbytta mot motsvarande synonymer. Tydligen ska uppemot 50 % av orden i en text kunna bytas ut mot synonymer, och texten kommer fortfarande inte undvika att bli ihopmatchad med sin originalkälla.

Ett annat enkelt sätt att ändra innehållet i en text är att flytta om ordningen på meningar eller bara fylla på med andra meningar mellan de redan existerande. Ett sätt att göra detta vore att ta två separata texter och kombinera dessa genom att

skapa en ny text där varannan mening kommer från de olika dokumenten. Enligt Turnitin så upptäcker deras verktyg detta. De betonar att detta försök att undkomma plagieringskontrollen ofta får motsatt effekt, då det i stället för en källa finns två *olika* källor att matcha texten med [11].

5.2 Sökmotorn Google

Google använder sig av tekniken PageRank för att undersöka webblänkstrukturer och bestämma vilka webbsidor som är viktigast när en specifik sökning har gjorts. En hypertextmatchningsanalys utförs sedan för att fastställa vilka webbsidor som är relevanta. På detta sätt kan Google visa de mest relevanta och pålitliga resultaten först.

Det maximala antalet tecken som sökmotorn Google kan ta som indata är 32 ord (2^5 stycken), vilket betyder att endast kortare stycken kan granskas åt gången. För den intresserade läsaren finns mer ingående information om PageRank, hypertextmatchning samt sökmotorer i allmänhet att läsa i (Manning, Raghavan & Schütze, 2008)[1] [3].

5.3 Övriga plagieringskontrollsverktyg

Utöver Turnitin och sökmotorn Google finns det ett flertal andra plagieringskontrollsverktyg som kan användas. Några av dessa är URKUND samt Plagiarism Detect system, som beskrivs kortfattat nedan.

URKUND

URKUND (urkund.se) är det ledande systemet för plagiatkontroll i Norden och används idag av majoriteten av högskolor och universitet i Sverige. URKUND jämför de dokument som skickas in med över 3000 förlagsdatabaser som bland annat innehåller e-böcker, databaser, tidningar och böcker som ej finns som e-böcker, men som de har tillgång till genom unikt samarbete med förlaget. Inskickade dokument jämförs även med det “öppna” Internet, vilket innebär att URKUND hittar alla matchningar som kan hittas med en vanlig sökmotor. Dessa dokument kommer även att jämföras med det “stängda” Internet, de delar av Internet vars material endast går att komma åt genom inloggning. Slutligen jämförs dokumenten även med URKUNDs egen databas över tidigare inskickat material, en databas som nu innehåller uppemot 4 miljoner uppsatser.

URKUND använder en egenutvecklad algoritm för att matcha de inskickade dokumenten, men denna algoritm är inspirerad av flera kända algoritmer. Då UR-

KUND anser att deras algoritm är den bästa av de förekommande är vidare information om denna hemlig.

När en uppsats skickas in till URKUND kontrolleras dels hela texten, men även delar av texten, ända ner till ordnivå. De gör nyckelordssökningar, strängsökningar och hela textsökningar. Med hjälp av synonymsökningar hittar verktyget texter som är kraftigt omskrivna. URKUND är inte strängberoende, vilket innebär att det inte handlar om hur många ord i följd som använts i en text, utan istället om den "poäng" ett stycke får i jämförelsen. Detta innebär att URKUND alltså kan hitta en förlaga till en text även om ordföljden är omkastad, orden är blandade eller om ord är utbytta mot synonymer. Förekommer nyckelord och ordkombinationer liknande en annan text flera gånger i en text, rankas träffen upp poängmässigt. Genom dessa nyckelordssökningar kan URKUND även hitta texter som är översatta från ett annat språk.

Viktigt att påpeka är att URKUND aldrig avgör vilka texter som är ett plagiat. De hittar och jämför texter och lämnar sedan underlaget till läraren och det är sedan lärarens uppgift att själv avgöra vad som har hänt i den kontrollerade uppsatsen [19].

Plagiarism Detect system

"Plagiarism Detect system" är ett plagieringskontrollsverktyg som finns på Internet. Det finns både en gratis- och en betalversion. Gratisversionen använder sig av Google API för att granska texterna. Idén går ut på att dela upp texten i meningar, dessa meningar omvandlas sedan till frågor som en åt gången används som indata till sökmotorn Google. Detta verktyg ger dock inte hundra procent noggrannhet och är endast begränsat till hundra ord. [10] Betalversionen däremot använder multi-layers teknologi tillsammans med olika APIs [9].

Del II

Undersökning med resultat och analys

Kapitel 6

Tillvägagångssätt

6.1 Metod

För att dra rimliga slutsatser om vad plagieringskontrollsverktygen klarade och inte klarade av, undersöktes varje specifikt plagieringskontrollverktyg separat. Genomtänkta testtexter utformades och redigerades under arbetsprocessens gång. Texterna manipulerades på olika sätt utifrån resultaten för att hitta plagieringskontrollsverktygens styrkor och svagheter.

Till att börja med gjordes generella undersökningar. Plagieringskontrollsverktygens databaser klarlades, detta för att få kunskap om vilka webbsidor, böcker och skrifter de faktiskt har tillgång till och på det viset kunna utnyttja den vetskapen. Generella tester, såsom vilka dokumentformat plagieringskontrollsverktygen klarar av, synonymbyte av ord, texter som är översatta från andra språk samt andra tester som kan vara intressanta, gjordes även.

En mer fördjupad undersökning följde sedan. Grundtexterna som användes var texter som överensstämmer till hundra procent med en källa, alltså hundra procent plagiering av någon annans litterära verk.

För att få struktur på testerna användes den svenska grammatikens regler och bestämmelser som grund. Testerna gick ut på att identifiera varje ords ordklass i texterna och på det viset systematisk byta ut alla ord i samma klass mot synonymer. Målet var att kartlägga vilka ordklasser som kunde bytas ut, enskilt eller i kombination, mot synonymer för att göra texten så pass oigenkännlig att plagieringskontrollsverktygen inte hittar den ursprungliga källan.

Möjligheten att ändra ordklassers form undersöktes även, genom att ändra formen på en mening blir texten svårare att känna igen för ett plagieringskontrollsverktyg.

6.2 Begränsningar

Då tiden att undersöka kapaciteten hos alla populära plagieringskontrollverktyg inte motsvarar den tid som detta projekt skall omfatta, så har vissa begränsningar gjorts. Då Kungliga Tekniska Högskolan (KTH) har tillhandhållit plagieringskontrollsverktyget Turnitin, så kommer fokus ligga på detta verktyg.

Utöver Turnitin kommer sökmotorn Google också att granskas och analyseras för att kunna upptäcka eventuella skillnader mellan verktygen. Då gratisversion av verktyget "Plagiarism Detect system" använder sig av Google API, så kommer inte detta verktyg att undersökas och på grund av ekonomiska skäl kommer inte heller betalversionen att prövas. URKUND har som policy att inte ge studenter tillgång till deras verktyg, vilket leder till att deras verktyg inte kan undersökas i denna rapport.

För att kunna ställa resultaten mot varandra och dra rimliga slutsatser, har valet gjorts att alla texter i undersökningarna ska vara skrivna i samma språk. Då uppsatsen är på svenska, föll det naturligt att använda svenska texter.

Den sista begränsningen är att bortse från den mänskliga faktorn, med det menas att undersökningen inte tar hänsyn till vad en lärare skulle tyckt om läraren läste texten som blivit manipulerad.

Kapitel 7

Generella Tester

Både Turnitin och Googles sökmotor kommer att sättas på prov och undersökas. Då plagieringskontrollsverktygen använder sig av egna databaser samt Internet är det intressant att undersöka hur mycket material de har tillgång till, samt huruvida de har tillgång till lösenordskyddade webbsidor eller webbplatser som innehåller gamla inskickade skoluppsatser eller ej.

7.1 Verktygens funktionalitet

En mycket generell undersökning har gjorts för att undersöka om testresultaten överensstämmer med vad varje enskilt plagieringskontrollsverktyg har utlovat användaren, beträffande funktionalitet och kapacitet.

7.1.1 Turnitin

Turnitin klarar av att hitta och matcha texter även om en mening inte är till hundra procent lik originalmeningen. Verktyget ger då en uppskattning av hur mycket av meningen som var lik, detta ges i procent. Om man byter plats på alla ord i en mening kommer Turnitin fortfarande kunna matcha till hundra procent.

7.1.2 Sökmotorn Google

Sökmotorn Google klarar av att hitta kortare stycken som är direkttagna från en Internetkälla, även om orden är omkastade. Däremot är det inte garanterat att ett stycke hittas om några ord blivit utbytta, tagits bort eller nya ord blivit tillagda. Googles sökmotor har mycket svårt för att hitta källan till en sökmening om något av orden i texten är felstavat. Det räcker med ett felstavat ord i en mening för att källan till denna ej skall hittas.

7.2 Databaser och Internet

Det är endast Turnitin som använder sig av lokala databaser. Att använda sig av egna databaser är en styrka, eftersom verktygen då har tillgång till alla gamla uppsatser som tidigare skickats in. Detta kan bekräftas då ett flertal redan inskickade uppsatser återinskickades till Turnitin och gav resultatet hundra procent matchning.

En undersökning angående tillgången till böcker har även gjorts. Utdrag från fysiska böcker har testats som indata till Turnitin och Google. Inget verktyg kunde matcha de tagna styckena. Samma resultat erhölls då tester bestående av kortare stycken från e-böcker och "Google böcker" gjordes, undantaget var dock att sökmotorn Google lyckades matcha stycken tagna från "Google böcker".

Resultatet av undersökningen av plagieringskontrollsverktygens tillgång till lösenordskyddade sidor [1] uppvisade att Turnitin inte hade åtkomst till dessa webbsidor. Eftersom det finns webbsidor [2] som uppmanar studenter att lägga upp sina uppsatser på deras sidor, kan det vara enkelt för andra studenter att använda sig av dessa. Då tester av uppsatser från dessa sidor gjordes visade det sig att Turnitin inte hade åtkomst till dessa och bedömde den tagna uppsatsen som icke–matchande. Detta var till skillnad från sökmotorn Google, som kunde lokalisera textens ursprung.

7.3 Översättning från andra språk

Då plagiering innefattar en direktöversättning från ett språk till ett annat har undersökning kring möjligheten att översätta en text till ett annat språk gjorts. Verktyget "Google översätt" som Google tillhandahåller har använts för att undersöka eventuella översättningar. Varken Turnitin eller sökmotorn Google lyckades känna igen en text som är översatt från ett språk till ett annat. Resultatet blev noll procent matchning från samtliga verktyg.

7.4 Synonymer

Då det är enkelt att byta ut enstaka ord i en text mot synonymer har tester kring detta gjorts. Resultat uppvisade att synonymutbyte drog ner procenten då texten granskades av Turnitin. Ju fler synonymer som byts ut, desto mer svårigenkännlig blir texten. Turnitin klarade av att till en viss punkt urskilja de resterande orden till rätt källa. Detsamma gällde sökmotorn Google. Då denna taktik visade sig

[1] exempel: Nationalencyklopedins webbsida
[2] mvgplus.se, mimersbrunn.se

användbar för att lura plagieringskontrollerna kommer en fördjupad redogörelse av denna metod att presenters i avsnittet 8.

7.5 Dokumentformat

För att Turnitin ska kunna granska ett dokument så kan inte dokumentet vara i vilket format som helst. En undersökning av de godkända formaten har gjorts för att kunna se om det har någon betydelse. Resultatet gav att Turnitin inte klarade av skyddade PDF-filer, vilket betyder att det inte går att redigera, markera i filen. Det framgår inte av Turnitin att dokumentet är skyddat, och Turnitin kommer därmed att klassa texten som icke–matchande, vare sig texten är plagierad eller inte.

Inga andra problem uppstod då de andra godkända dokumentformaten undersöktes.

Kapitel 8

Fördjupning av tester

8.1 Turnitin

Då det visade sig att synonymbytning gav önskat resultat, gjordes en fördjupad undersökning. Även tester då stycken byter plats med varandra, meningsuppbyggnaderna förändras samt nya texter skapas av meningar från olika texter har undersökts.

8.1.1 Byta plats på stycken

En undersökning har gjorts på texter som varit mellan 1–2 A4 (Times New Roman, storlek 12) sidor långa. Varje enskild text delades in i stycken, dessa stycken kastades sedan om så att de hamnade i en ny ordningsföljd.

Resultat

Resultatet från undersökningen presenteras i tabell 8.1. Tabellen beskriver vad som sker då varje stycke kortas ner och om det är någon skillnad beroende på hur långt stycket är.

Längd på stycket *antal rader**	**Resultat** *Turnitins uppskattning av matchningsprocenten*
4–6	100
2–3	100
1-2	98–99

*Räknat då textfonten är Times New Roman, storlek 12 och styckena är tagna från ett doc-document.

Tabell 8.1. Visar medelvärdet av resultatet från tester då stycken bytte plats i en text.

Analys

Från resultatet kan slutsatsen dras att det inte spelar någon roll om styckena byter plats. Bryter man ner texten till endast meningar och sedan blandar om dessa, så kommer verktyget Turnitin ändå känna igen texten till nästan hundra procent.

8.1.2 Kombinera olika texter

För att undersöka hur bra plagieringskontrollsverktyget är på att matcha en text som härstammar från flera olika texter, gjordes tester där meningar från ett antal olika texter klipptes ihop till en enda stor text. Alla texter som använts är texter som hittas av Turnitin. Det nya dokumentet bestod först av en del från text 1, sedan från text 2, text 3, och så vidare ett antal gånger om. Först utfördes detta med fem meningar på raken från en och samma text, sedan med tre meningar på raken och till slut med endast en mening i taget från varje text. Tester utfördes först med fem olika texter och sedan med tio olika texter i samma dokument.

Resultat

Att kombinera fem meningar på raken från en och samma text och sedan upprepa detta med alla de fem texterna gång på gång, är inget som påverkar huruvida plagieringskontrollsverktyget hittar ursprungskällorna eller ej. En sådan text får 100 % i matchningsprocent, då plagieringskontrollsverktyget inte har några problem att matcha alla dessa stycken. Detta gäller även i fallet där tre meningar i rad kommer från en och samma text. Om det däremot endast är en mening i taget från en och samma källa, så påverkar detta resultatet något och matchningsprocenten blir 96 % istället. Resultaten framgår i tabell 8.2.

Om antalet texter utökas till tio stycken istället och det testdokument som används har en mening i taget från vardera av dessa tio texter, sjunker matchningsprocenten till 85 %. Då detta fortfarande är en hög siffra har inga ytterligare tester gjorts på detta specifika fall.

Antal kombinerade texter	**Antal meningar i rad**	**Resultat** *Turnitins uppskattning av matchningsprocenten*
5	5	100
5	3	100
5	1	96
10	1	85

Tabell 8.2. Resultatet i matchningsprocent då flera texter klipptes ihop till en

Analys

Precis som Turnitin själva säger, så upptäcker deras plagieringskontrollsverktyg att texten kommer från fem olika källor och markerar alla dessa med varsin färg. Testdokumenten fick en väldigt hög matchningsprocent på omkring 90 %, med andra ord i stort sett hela texten. De delar av texten som inte matchades var sådana meningar som var mycket allmänna eller vanliga och inte ensamstående går att matcha till en specifik källa då de används på alltför många ställen.

Att ta ett extremt stort antal texter, exempelvis 100 stycken olika texter och kombinera dessa på liknande vis som ovan, gör visserligen att det finns en chans att matchningsprocenten sjunker. Däremot så kommer plagieringskontrollen förmodligen att lyckas matcha merparten av texten och upptäcka det (baserat på uppvisade testresultat). Tidigare i rapporten nämndes att ett större antal källor även ger plagieringskontrollsverktyget ett större antal källor att matcha den inskickade texten med. Med andra ord är det ingen större idé att försöka sig på att lura plagieringskontrollen på detta vis.

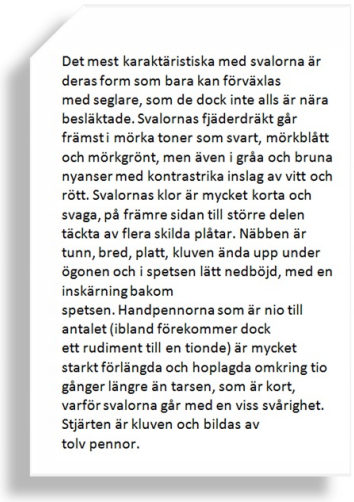
8.1.3 Lingvistiska modifieringar

Då det framkom i de generella testerna i avsnitt 7.4 att manipulation av ord var ett effektivt sätt för att få ner procenten då texten granskades av Turnitin, så kommer en djupare undersökning om detta att följa. Målet med undersökningen är att den modifierade texten ska vara så pass oigenkännlig att Turnitin inte känner igen textstycket. Om man då med hjälp av synonymer ska byta ut särskilda ord i en mening måste detta göras på ett mycket genomtänkt sätt. Alla ord i en mening bör lämpligtvis inte bytas ut mot synonymer, risken att meningen skulle få en annan innebörd är då stor. De ord som har en avgörande roll i en text, så att läsaren ska kunna förstå vad texten handlar om, är vanligtvis ord tillhörande de öppna ordklasserna. I undersökningen kommer därför endast ord tillhörande öppna ordklasser att undersökas.

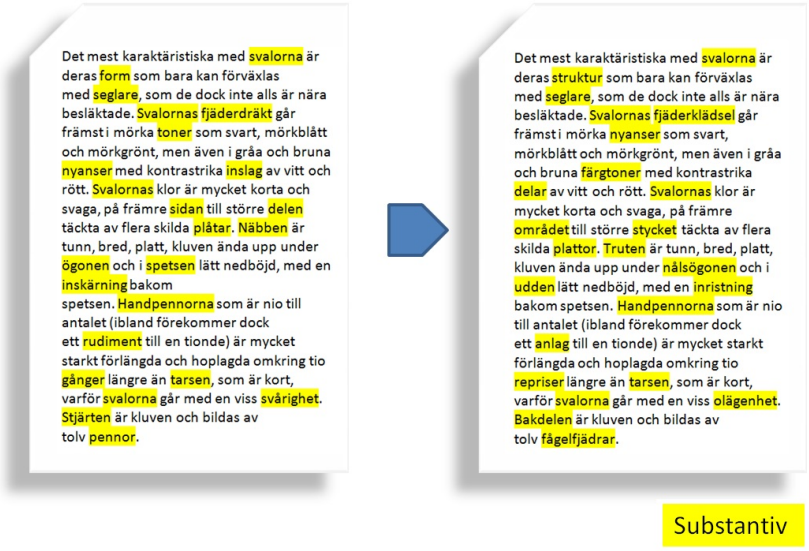
Resultat

En ordklass i taget har undersökts. Alla ord tillhörande en specifik öppen ordklass har då blivit utbytta mot synonymer i varje text som ska modifieras. Varje text som används i undersökningen omfattar ett A4-ark (Times New Roman, storlek 12) och texterna kommer från wikipedia.se, vilket betyder att de kommer hittas av Turnitin. De ordklasser som undersökts är: substantiv (NN), verb (VB), adjektiv (JJ) och adverb (AB). Alla ord har blivit utbytta mot synonymer som synonymer.se har föreslagit.

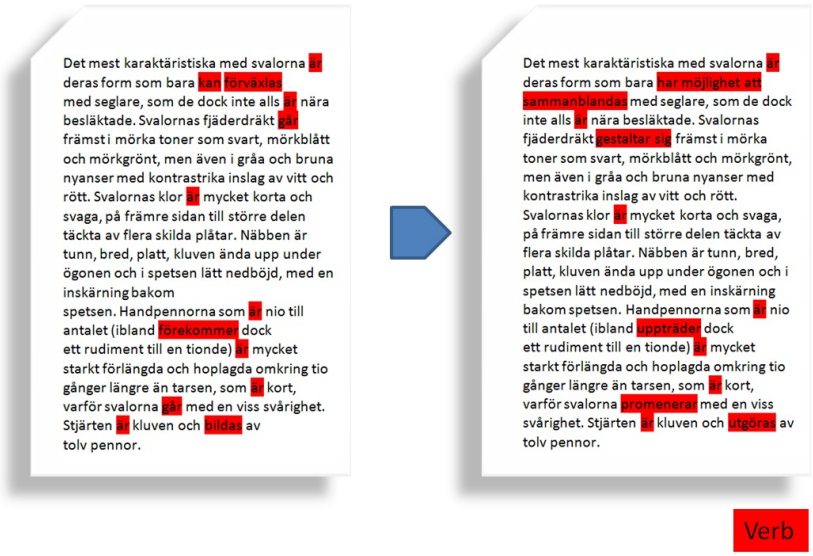
Figurerna 8.2–8.5 illusterar hur det kan se ut när alla ord tillhörande en viss ordklass blivit utbytta mot synonymer. Hur texten ser ut från början kan ses i figur 8.1.



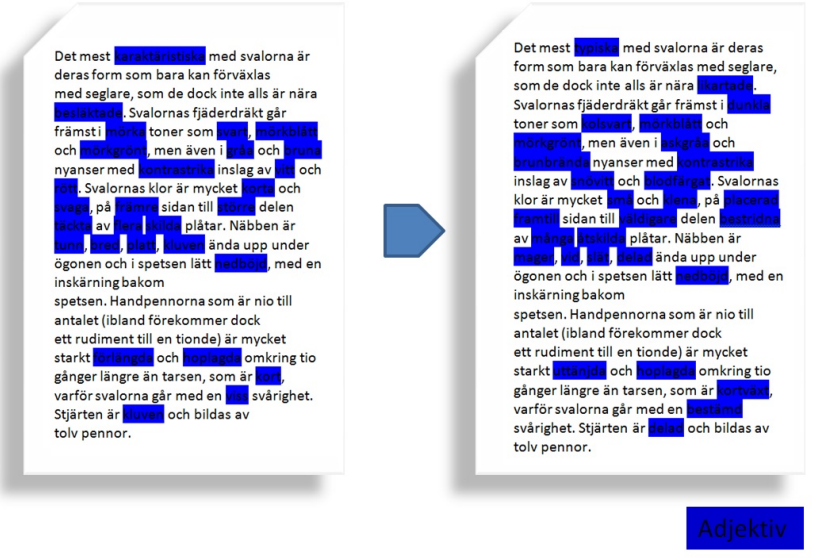
Figur 8.1. Ett exempel på hur en text kan se ut innan den manipuleras. Denna text används i alla figurexempel i detta kapitel. Texten kommer från wikipedia.se [18].



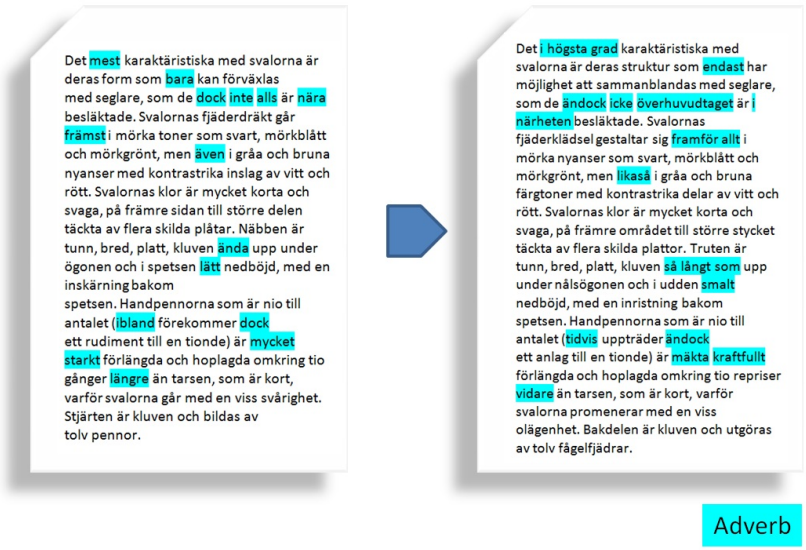
Figur 8.2. Alla ord som tillhör ordklassen substantiv har blivit utbytta mot synonymer. Orden som är markedade i gult är substantiv. Till vänster ses originaltexten och till höger den manipulerade texten.



Figur 8.3. Alla ord som tillhör ordklassen verb har blivit utbytta mot synonymer. Undantaget är ordet"är", på grund av att det inte finns någon synonym som motsvarar ordet"är". Orden som är markedade i rött är verb. Till vänster ses originaltexten och till höger den manipulerade texten.



Figur 8.4. Alla ord som tillhör ordklassen adjektiv har blivit utbytta mot synonymer. Orden som är markedade i mörkblått är adjektiv. Till vänster ses originaltexten och till höger den manipulerade texten.



Figur 8.5. Alla ord som tillhör ordklassen adverb har blivit utbytta mot synonymer. Orden som är markedade i ljusblått är adverb. Till vänster ses originaltexten och till höger den manipulerade texten.

Ordklass	**Turnitins bedömning** *Hur stor andel av texten som är plagierad enligt Turnitin (%)*	**Plagierat** *Hur stor andel av texten som faktiskt är plagierad (%)*
NN	82	84
VB	88	89
JJ	92	94
AB	90	92

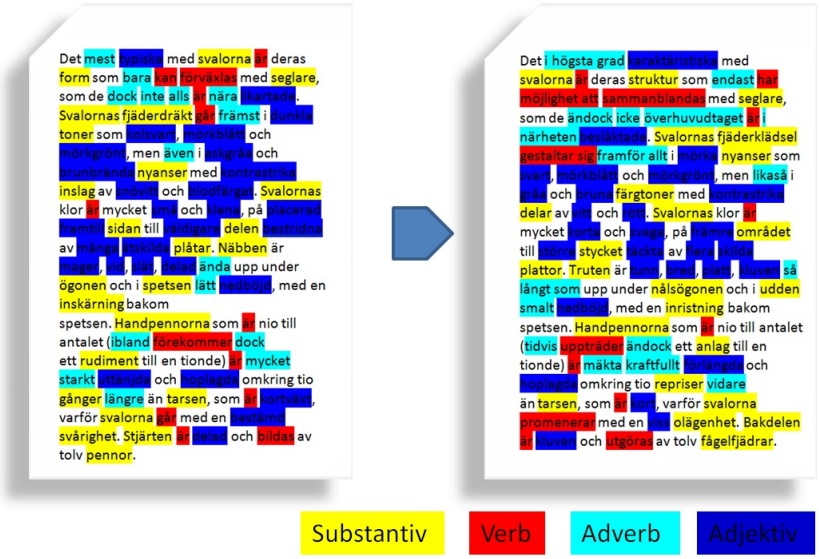
Tabell 8.3. Medianvärdet av de procentvärden som Turnitin gav då verktyget granskade texterna då ord hade byts ut mot synonymer.

Resultatet som erhölls då 12 texter manipulerades enligt ovanstående riktlinjer och sedan granskats av Turnitin kan ses i tabell 8.3.

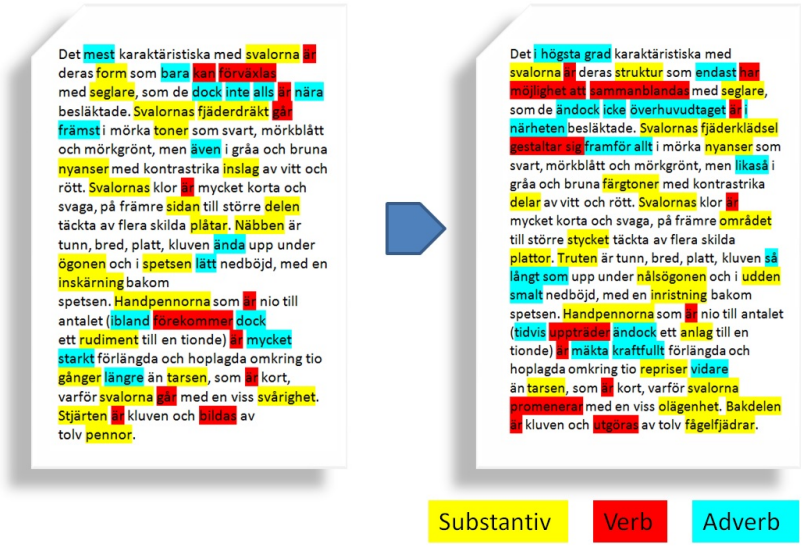
Vidare gjordes två undersökningar då först alla ord som tillhör någon av de fyra ordklasser i en och samma text byttes ut mot synonymer. Figur 8.6 illusterar hur det kan se ut när alla ord tillhörande NN, VB, JJ och AB blivit utbytta mot synonymer.

På grund av resultatet som ges i tabell 8.3 har även en undersökning gjorts, angående möjligheten att bortse från att byta ut de ord som tillhör en ordklass som gav sämst resultat. Ordklassen JJ gav sämst resultat och bortsågs därför. Detta test gjordes för att komma närmare den brytpunkt då Turnitin inte längre kan

identifiera texten. Resultatet presenteras i tabell 8.4. Hur en text kan se ut om ord tillhörande flera olika ordklasser bytts ut kan ses i figur 8.6 och 8.7.



Figur 8.6. Alla ord som tillhör ordklasserna NN, AB, JJ och AB har blivit utbytta mot synonymer. Orden som är markedade i färger tillhör en specifik ordklass. Till vänster ses originaltexten och till höger den manipulerade texten.



Figur 8.7. Alla ord som tillhör ordklasserna NN, AB och AB har blivit utbytta mot synonymer. Orden som är markedade i färger tillhör en specifik ordklass. Till vänster ses originaltexten och till höger den manipulerade texten.

Ordklass	**Turnitins bedömning (Median)** *Hur stor andel av texten som är plagierad enligt Turnitin (%)*	**Plagierat** *Hur stor andel av texten som faktiskt är plagierad (%)*
NN+VB+AB	0	62
NN+VB+JJ+AB	0	57

Tabell 8.4. Medianvärdet av de procentvärden som Turnitin gav då verktyget granskade texterna då NN, VB, JJ och AB, samt när NN, VB och AB byttes ut. Samt det riktiga procentvärdet av ord som var utbytta.

För att kunna avgöra vilka ordklasser som påverkar Turnitins bedömning mest, så jämfördes den faktiska andelen av utbytta ord mot Turnitins resultat i matchningsprocent. Resultatet kan ses i tabell 8.5 och 8.6.

Text	**NN**	**VB**	**JJ**	**AB**	**NN+VB+AB**	**NN+VB+JJ+AB**
1	83	88	93	90	63	57
2	87	79	95	92	60	55
3	81	93	92	95	70	62
4	83	82	96	95	61	57
5	83	89	92	88	61	53
6	85	89	95	91	62	58
7	82	84	90	91	57	49
8	83	89	94	93	62	55
9	85	86	93	92	65	58

Tabell 8.5. Hur stor andel av texten som faktiskt är plagierad. I (%).

Text	**NN**	**VB**	**JJ**	**AB**	**NN+VB+AB**	**NN+VB+JJ+AB**
1	83	85	93	90	0	0
2	86	78	93	91	0	0
3	80	93	92	95	68	52
4	0	0	95	95	0	0
5	83	89	92	88	56	6
6	46	86	95	91	0	0
7	0	84	90	91	0	0
8	83	89	94	91	62	20
9	85	86	93	91	60	44

Tabell 8.6. Hur stor andel av texten som är plagierad enligt Turnitin. I (%).

Analys

Den första slutsatsen som kan dras är att det skiljde sig beroende på vilken ordklass som bytts ut, vilket kan ses i tabell 8.3. Turnitin hade svårare att kunna identifiera en text då substativen byttes ut mot synonymer jämfört med om adjektiven hade bytts ut. Oftare är det enklare att byta ut ett substantiv mot ett annat om man ska se från en grammatisk synvinkel. Adjektiv kan ofta vara så specifika att det är svårt att hitta en synonym som motsvarar originalordet, exempel på detta är färger och mått. Helst skulle man vilja byta ut så få ord som möjligt för att minimera risken att grammatiska fel ska uppstå i texten. Målet är att hitta brytpunkten, det vill säga hur stor del av texten som måste bytas ut för att Turnitin inte längre ska känna igen den manipulerade texten. I tabell 8.4 ses resultatet då just detta undersöktes, möjligheten att kombinera olika ordklasser för att undkomma att Turnitin klassar texten som plagiat. Dock spelar det ingen större roll om det var tre eller fyra ordklasser som byttes ut, detta kan ses i tabell 8.6. Antingen kände Turnitin igen texterna för båda eller för ingen. Intressant att poängtera är att i vissa texter, så räckte det med att byta ut alla substantiv eller verb för att texten skulle bli oigenkännlig för Turnitin.

Utifrån denna undersökning kan slutsatsen dras att alla tre eller fyra testade ordklasser måste bytas ut för att lura Turnitin. Det är högre sannolikhet att undkomma Turnitin om fyra ordklasser byts ut istället för tre. Resultaten pekar på att det dessutom krävs att texten till stor del består av substantiv och verb för att texten ska skall bli oigenkännlig för Turnitin. I tabell 8.7 ses resultatet då utbyte av substantiv och verb kombinerats och huruvida Turnitin därmed lyckades känna igen texten eller ej.

Text	**Turnitins bedömning** *Andel av texten som är plagierad enligt Turnitin då NN+VB blivit utbytta (%)*	**Plagierat** *Andel av texten som faktiskt är plagierad då NN+VB blivit utbytta (%)*	**Lyckades Turnitin matcha texten?** *NN+VB+JJ+AB har blivit utbytta*
1	84	85	Nej
2	82	83	Nej
3	87	87	Ja
4	0	83	Nej
5	86	86	Ja
6	66	87	Nej
7	42	83	Nej
8	86	86	Ja
9	86	86	Ja

Tabell 8.7. Medelvärdena då NN och VB har kombinerats från tabell 8.6, samt huruvida Turnitin lyckas känna igen texten eller ej.

Slutsatsen som kan dras utifrån tabell 8.7 är att det maximalt får vara 86 % kvar av originaltexten efter att substantiv plus verb har bytts ut. Då är chansen stor att Turnitin inte kan känna igen texten, om även adjektiven och adverben också byts ut. Alltså måste texten innehålla minst 14 % substantiv och verb, eftersom 100 % - 86 % = 14 %.

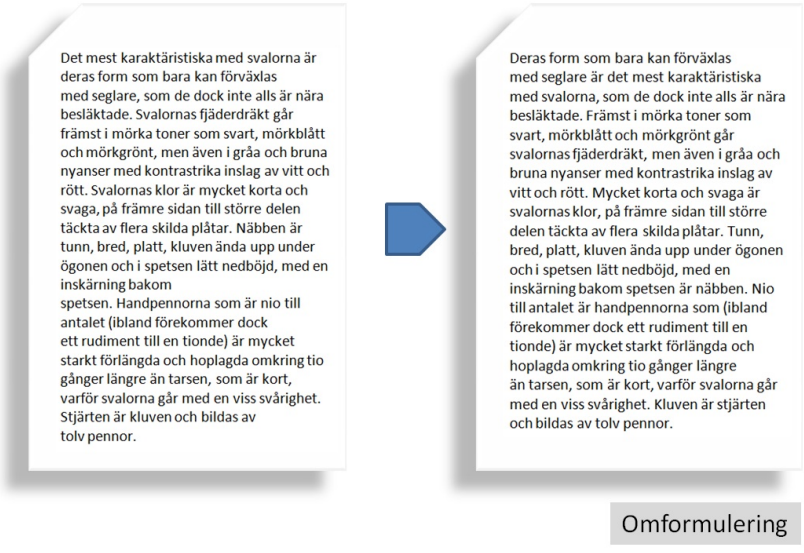
8.1.4 Permuterande modifieringar

Möjligheten att modifiera hela meningar har även undersöks. En bra metod för att ändra en meningsuppbyggnad är att gå från passiv form till aktiv eller tvärtom. Även lättare omformuleringar går att göra i vissa fall, beroende på vilka ordklasser som ligger i följd. Specifika regler kan definieras gällande hur olika meningsfall kan formuleras om för att fortfarande vara grammatiskt korrekta.

En undersökning gjordes för att se hur Turnitin skulle bedöma texterna då meningarna omformulerades. Varje text som används i undersökningen omfattar ett A4–ark och texterna kommer från wikipedia.se, vilket betyder att de kommer hittas av Turnitin.

Resultat

Hur en text kan se ut efter att meningarna i texten blivit omformulerade kan ses i figur 8.8.



Figur 8.8. Ett exempel på hur en text kan se ut efter att vissa omformuleringar av meningar har gjorts. Till vänster ses originaltexten och till höger den manipulerade texten.

Resultatet som erhölls då 14 texter modifierades enligt ovanstående riktlinjer och sedan granskats av Turnitin kan ses i tabell 8.8

Metod	**Turnitins bedömning(median)** *Hur mycket som är plagierat enligt Turnitin(%)*
Omformulering	83

Tabell 8.8. Median av de procentvärden som Turnitin gav då verktyget granskade texterna då omformuleringar hade gjorts.

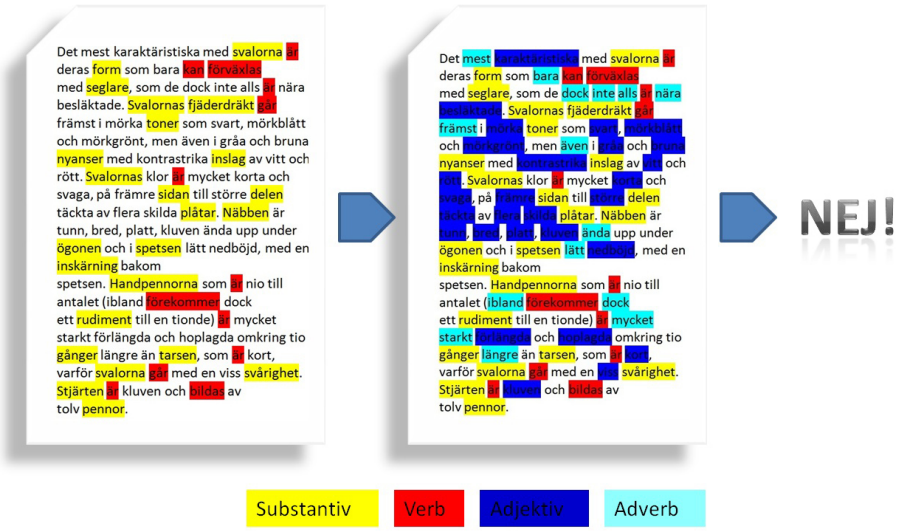
För att sänka procenten ytterligare så har en undersökning utförts då synonymbytning kombinerats med omformulering av meningar. Undersökningen presenteras i tabell 8.9.

Ämne	**Omformulering (%)** *Hur stor andel av texten som är plagierad enligt Turnitin (%)*	**Omformulering +NN+VB+AB(%)**	**Omformulering +NN+VB+JJ+AB(%)**
1	80	0	0
2	81	0	0
3	82	38	25
4	80	0	0
5	83	35	0
6	82	0	0
7	90	0	0
8	84	40	13
9	72	30	21

Tabell 8.9. Visar resultatet för de olika texter som manipulerades då texterna omformulerades och synonymer byttes ut.

Analys

Som återspeglas i tabell 8.9, upptäcktes fyra av nio texter då substantiv, verb och adverb bytts ut samt omformulering gjorts. Då alla fyra ordklasser byttes ut i kombination med omformulering, upptäcktes istället tre av nio texter, vilket är ett bättre resultat. Detta innebär att en elev i regel skulle kunna anklagas för plagiering var tredje gång, om alla dessa omskrivningar gjorts. Med andra ord, så skulle en elev i två av tre fall kunna komma undan med detta. En illustration om hur detta skulle gå till ges i figur 8.9. För att till hundra procent inte kunna bli anklagad för plagiering, skulle ytterligare modifikation av varje text behöva göras.



Figur 8.9. Om alla ord som tillhör ordklasserna substantiv och verb markeras. Om dessa markerade ord utgör över 14 % av hela texten så markeras även ord som tillhör ordklasserna adjektiv och adverb. Om alla dessa markerade ord byts ut mot synonymer plus att meningarna formuleras om, så är chansen stor att Turnitin inte känner igen texten längre.

8.2 Sökmotorn Google

I de följande testerna har enstaka meningar från ett antal olika källor använts, där dessa källor med säkerhet hittas genom en sökning på Google. Dessa meningar har sedan ändrats och omstrukturerats på olika sätt för att undersöka om sökmotorn fortfarande kan relatera dessa meningar till deras ursprungliga källa. Samma meningar har använts i alla tester och alla tester har utförts på samtliga av dessa meningar. Varje test har utförts på en mening i taget.

8.2.1 Borttagning av ord

För att undersöka hur viktigt det är att alla ord från en originalmening finns med som sökord, gjordes tester där utvalda ord togs bort från meningarna. Systematiskt togs först det första ordet i meningarna bort, följt av det andra, tredje och så vidare fram till det sista ordet i meningen. Endast ett ord i taget togs bort i detta fall. Därefter undersöktes resultaten av att ta bort fler ord än ett i taget från originalmeningen. Två ord togs bort, därefter tre ord och så vidare ner till dess att endast tre av orden var kvar. Detta utfördes gång på gång för att undersöka om det fanns någon skillnad mellan resultaten beroende på vilka av orden som behölls som sökord.

Resultat

Googles sökmotor klarar av att hitta ursprunget till en mening även om denna mening ej är komplett. Att ta bort ett enstaka ord påverkar inte sökresultatet över huvud taget, då sökmotorn fortfarande hittar källan utan problem. Detta gäller även då en större mängd ord tas bort från meningen, så länge som de kvarvarande orden inte är alltför generella. Det spelar alltså roll vilka ord som används i sökningen och vilka ord som tas bort.

Analys

Generellt visar testerna att det främst är ord som tillhör öppna ordklasser som är viktiga att ha kvar som sökord, då dessa ord är mer specifika och bestämmande. Ord från slutna ordklasser har en tendens att vara alldeles för allmänna för att vara avgörande i denna typ av sökning. Ofta finns det i en text ett eller flera kärnord som är viktiga och grundläggande begrepp för just det specifika ämne som texten behandlar. Detta är sådana ord som är viktiga att ha kvar som sökord i en mening.

8.2.2 Byta plats på ord

För att undersöka om ordningsföljden på orden i en mening är relevant för att Googles sökmotor ska hitta en källa, gjordes en mängd olika omstruktureringar av orden i enskilda meningar från separata källor. De nybildade meningarna blev inte alltid grammatiskt korrekta, men innehöll alltid samma ord som originalmeningen, dock i olika följder. Orden skrevs in på följande olika sätt:

- Baklänges
- Vartannat ord från början och vartannat ord från slutet av meningen
- Först alla ord med udda position, följt av orden med jämn position
- Slumpmässigt

Resultat

Att ta en mening och sedan flytta runt orden i meningen hur som helst lurar ej Googles sökmotor. Om den totala mängden av sökord är densamma som i originalmeningen, så kommer sökmotorn att matcha dessa ord och genom detta hitta originalkällan ändå. Detta gäller oavsett i vilken ordning orden kommer och vare sig meningen är grammatiskt korrekt eller ej.

Analys

Då det inte spelar någon roll i vilken ordning orden kommer, är detta ej en rekommenderad metod att använda för att plagiera en text. På grund av detta kommer

denna typ av modifiering av texter ej att prövas vidare i undersökningen av sökmotorn Google.

8.2.3 Byta ut ord mot synonymer

För att undersöka om det skiljer sig mellan olika ordklasser, vilka ord som används som sökord, gjordes tester på detta. Mening för mening byttes ett ord i taget ut mot en synonym eller ett motsvarande ord. Därefter undersöktes om detta påverkade huruvida sökmotorn hittade originaltexten eller ej. Resultaten delades in efter ordklass och de ordklasser som undersöktes var substantiv, verb, adjektiv samt adverb.

Resultat

Vilken ordklass ett ord tillhör är inte direkt avgörande för hur viktigt ordet är som sökord. Det är ingen specifik ordklass där ett utbyte av ett sådant ord ger dramatiska konsekvenser. Vilken position i meningen ordet har är inte heller direkt avgörande. Däremot finns det ord i meningar, framförallt substantiv, men även verb som har stor vikt för att källan till meningen skall hittas av sökmotorn, se resultatvärden i tabell 8.10. Dessa ord är, som beskrivs i sektion 8.2.1, de ord som är viktigast i innehållet och som definierar texten. I allmänhet är det mycket svårare för sökmotorn att hitta ursprunget till en mening om ett av orden är utbytt. I knappt fyra fall av tio hittar sökmotorn källan till meningen.

Ordklass	**Andel hittade** *Anger i hur stor del av alla testfall som en mening hittades, förutsatt att ord av den angivna ordklassen var utbytt(%)*
NN	28
VB	26
JJ	56
AB	38
Totalt	37

Tabell 8.10. Visar hur stor andel av testfallen där källan hittades, när ord byttes ut mot synonymer.

Analys

Då kärnan i många texter i själva verket är substantiv, är det logiskt att det är dessa ord är extra viktiga att ha kvar som sökord. Däremot är det en aning förbryllande att det i många fall räcker med ett enda felaktigt ord i en sökmening för att källan inte skall hittas.

Del III

Att “lura” Turnitin

Kapitel 9

Teori

Tillvägagångssättet för att "lura" Turnitin bygger på att alla ord i en text identifieras, det vill säga att alla ord tilldelas endast en ordklass. Ordklasstaggning kan användas till just detta. Då vissa ord kan tillhöra fler än en ordklass, krävs viss förkunskap om ordklasstaggning för att läsaren ska förstå hur det är möjligt att utan mänsklig inverkan tilldela den rätta ordklassen.

9.1 Ordklasstaggning

Ordklasstaggning (även kallat taggning) är processen att tilldela ordklasser eller andra lexikala klasstaggningar till ord. Taggningen kan göras på isolerade ord såväl som på hela meningar och texter. Om ett isolerat ord ska taggas som är homonymt eller homograft, går det inte att veta vilken betydelse ordet är menat att ha av skribenten. I en mening eller text däremot så kommer alla ord bero på varandra, vilket betyder att det med stor sannolikhet går att avgöra varje ords rätta innebörd i sammanhanget. Varje ord kommer på det viset endast tillhöra en ordklass.

Algoritmen som ska utföra ordklasstaggning måste kunna avgöra vilken betydelse ett ordet har i varje sammanhang.

Ordklasstaggningsalgoritmer som kan användas vid taggning av texter faller inom två klasser: regelbaserad disambiguering och probabilistisk disambiguering. Utöver dessa två finns det transformationsbaserad disambiguering som är en kombination utav ovanstående.

9.1.1 Regelbaserad disambiguering

Med regelbaserad disambiguering menas att ord som har flera betydelser (taggar) disambigueras med hjälp av regler som tar hänsyn till sammanhanget. Det första

steget är att varje ord tilldelas en lista med potentiella ordklasser. Andra steget är att använda sig av listor av handskrivna disambigueringsregler som gör det möjligt att minska antalet ordklasser till endast en ordklass per ord. Exempel på en regel är att ett ord som följs av ett substantiv och förgås av en artikel taggas som ett adjektiv. Detta betyder att ju fler handskrivna regler desto bättre bör resultat bli [5].

9.1.2 Probabilistisk disambiguering

Probabilistisk disambiguering går ut på att använda statistisk för att taggdisambiguera meningar. Den modell som oftast används är N–te ordningens markovmodell där N antingen är 1 eller 2. Då man har uppskattningarna av sannolikheterna kan taggningen beräknas med dynamisk programmering genom att använda sig utav Viterbis algoritm.

9.1.3 Okända ord

Både regelbaserad disambiguering och probabilistisk disambiguering kräver ett lexikon som listar ord. Trots att det finns stora lexikon, så kommer det alltid finnas ord som inte är med i dessa lexikon. Ord som ska taggas kan dessutom vara felstavade, vilket också kommer göra det problematiskt vid taggningen eftersom orden då förmodligen inte kommer att finnas med i lexikonet. Det finns tre metoder som kan användas då ordet är okänt. Den första metoden går ut på att gissa på alla taggar som finns och låta disambigueraren välja den lämpligaste taggningen. Den andra metoden går ut på att undersöka om ordet kan bildas som en komposition av kända ord. Om ordet skulle vara en sammansättning, så taggas ordet på samma sätt som efterledet. Den tredje metoden analyserar hur kända ord taggas som slutar på samma bokstäver som det okända. Därefter taggas det okända ordet likadant [4].

Kapitel 10

Tillvägagångssätt för att "lura" Turnitin

Denna del kommer utifrån de resultat och analyser som resten av uppsatsen gett, beskriva hur ett verktyg för att "lura" Turnitin skulle kunna fungera. Då Turnitin är det enda bastanta verktyget som undersökts i denna rapport, är följande fiktiva verktyg utformat utifrån de resultat som framkom i undersökningen. Detta verktyg kan ej med hundraprocentig säkerhet garantera att Turnitin ej genomskådar modifieringarna. Då verktyget är utformat efter Turnitin är det heller ej möjligt att garantera att detta verktyg skulle kunna "lura" andra plagieringskontrollsverktyg.

För att en text som från början är 100 % plagiering ska kunna passera obemärkt genom plagieringskontrollen har fem åtgärder identifierats:

- **Steg 1:** Användaren skyddar viktiga/betydelsefulla ord
- **Steg 2:** Använda handskrivna regler
- **Steg 3:** Tagga alla ord
- **Steg 4:** Byt ut ord mot synonymer
- **Steg 5:** Aktiv/passiv form + omformuleringar av meningar

En text skall alltså genomgå alla dessa steg, men detta görs automatiskt av verktyget. Det enda steget där användaren påverkar är i steg 1, därefter sköter verktyget resten.

Steg 1 I vissa texter finns det ord som ej kan bytas ut, utan att förändra kontexten. Dessa ord har användaren möjlighet att skydda (flagga). Detta innebär att dessa ord kommer att ignoreras av verktyget för att säkerställa att dessa ord ej byts ut. Om användaren skickar in en text till verktyget och anser att utdatan blev skev till en följd av att vissa specifika ord felaktigt byttes ut, är det i steg 1 som användaren får chans att påverka resultatet. Om användaren alltså inte blev nöjd

med resultatet, kan fler ord skyddas och texten skickas in på nytt.

Steg 2 Det andra steget är att låta texten modifieras utifrån ett antal regler som är handskrivna av en lingvist. Dessa regler skall användas för:

- Specialfall av de ord som är svåra att byta ut med hjälp av synonymer
- Förkortningar som skall skrivas ut

  Ex: t.ex. –> till exempel
- De fall där flera ord kan skrivas om till ett ord

  Ex: till exempel –> exempelvis
- Skala bort överflödiga ord

Ord som byts ut enligt de första tre av ovanstående punkter, kommer sedan att flaggas för att indikera att dessa ord redan är utbytta och ej skall bytas ut igen. I dagens läge finns det en betaversion av ett program som klarar av att utföra just denna typ av modifieringar, med hjälp av separata filer där de handskrivna reglerna finns definierade. Detta program heter Rephrase, men klarar ej av att flagga ord.

Steg 3 Nästa steg är att alla ord i texten taggas efter vilken ordklass de tillhör. Detta görs med antingen regelbaserad disambiguering eller probabilistisk disambiguering, alternativt med hjälp av en kombination av de båda. Ett verktyg som finns tillgängligt idag och gör just detta är den taggdisambiguerare som skapades i projektet Granska, där Viggo Kann från Nada, KTH är ansvarig. Verktyget finns tillgängligt på `http://csc.kth.se/tcs/projects/granska/tagga.html`.

Steg 4 När alla ord i texten är taggade efter ordklass skall möjligheten att byta ut ord mot synonymer användas. Då homonymer och homografer förekommer relativt ofta i det svenska språket används ordklasstaggning här med fördel för att enklare identifiera vilket ord det rör sig om i de specifika fallen. De ordklasser som i de fall där det är möjligt, byts ut mot synonymer är substantiv, verb, adverb och eventuellt adjektiv. Eftersom ett ord kan ha en synonym som består av flera delord, så kommer de nya orden som införs att taggas om på nytt. Synonymerna hämtas från ett synonymlexikon och för att uppnå så relevanta synonymutbyten som möjligt, kommer en lingvist att ha gjort ett stort antal tester gällande vilka synonymer som används ofta och vilka som ej gör det. Detta ger statistik som sedan används för att i den mån det är möjligt uppnå ett så bra synonymutbyte som möjligt.

Steg 5 Till sist skrivs de meningar om, som är möjliga att skriva om enligt aktiv/passiv form, och med hjälp av kunskapen om vilken ordklass varje ord tillhör, kan meningarna omformuleras och skrivas om ytterligare.

Del IV

Analys

Kapitel 11

Diskussion

11.1 Reflektion och diskussion

Efter att alla tester gjorts och varje enskilt resultat från de fördjupade testerna analyserats, kan vissa skillnader ses gentemot det som Turnitin påstod sig klara av. I avsnitt 5.1 framgick att Turnitin påstår sig klara av dels att genomskåda när flera texter kombinerats till en enda, men även när ord bytts ut mot synonymer. Då flera texter kombinerades, visade resultaten att Turnitins påstående var korrekt, nästintill hela texten matchades med sina ursprungskällor. Däremot lyckades Turnitin ej genomskåda när ord byttes ut mot synonymer, vilket de påstod sig göra. Detta gäller texter skrivna på svenska, och därmed kan det ej bekräftas att detta även gäller för synonymutbyten i texter skrivna på exempelvis engelska.

Det visade sig att Turnitin var betydligt bättre på att känna igen texter, även när de var grovt modifierade, än sökmotorn Google. Däremot hittade Googles sökmotor vissa källor som Turnitin ej hittade. Då dessa två verktyg kompletterar varandra mycket bra, vore det ultimata ett verktyg som är en kombination av de båda. I dagsläget finns dock inget sådant verktyg. Istället kan exempelvis en lärare först använda Turnitin, för att sedan komplettera de icke–matchade meningarna som sökningar med hjälp av Googles sökmotor. Då mängden av sådana meningar kan komma att bli mycket stor, är det mest realistiska att läraren i detta fall endast utför denna Google–sökning med de meningar som verkar suspekta, det vill säga misstänkta plagiat.

Viktigt att poängtera är att resultatet då texter blivit modifierade kan tyckas låta något märkliga för det mänskliga örat, även om de är grammatiskt korrekta. Så kan även bli fallet då det fiktiva verktyg som beskrevs i sektion 10 används för att modifiera en text. Detta beror på att syftet med undersökningen var att eventuellt "lura" ett plagieringskontrollsverktyg, och ej att lura en lärare. Därmed har den

mänskliga ifrågasättningsfaktorn ej tagits i beaktning.

11.2 Förslag på förbättringar av Turnitin

Det finns några klara brister som Turnitin borde se över. Verktyget i sig fungerar väldigt bra över lag, dock är deras tillgång till den information som finns tillgänglig på Internet inte lika imponerande. Eftersom det är så pass lätt att få information från Internet idag, är det extra viktigt att plageringskontrollsverktyg har tillgång till både det "öppna" och "stängda" Internet. Det framkommer i resultatet att Turnitin inte har tillgång till svenska webbplatser som tillhör det "stängda" Internet. Turnitin är som sagt ett verktyg som används internationellt. Det kan då tänkas vara för omfattande att få tillgång till hela "stängda" Internet. Något som Turnitin skulle vinna på är om de kunde få tillgång till "Google böcker". De skulle då ha möjlighet att jämföra inskickade uppsatser mot böcker som de förmodligen ej har tillgång till annars.

Google tillhandahåller även ett annat bra verktyg, som undersöktes i avsnitt 7.3, närmare bestämt "Google Översätt". Då det i resultatet visade sig att Turnitin inte klarade av att känna igen en text som var direktöversatt, skulle ett samarbete med "Google Översätt" vara ett alternativ. En annan valmöjlighet är att använda ett annat översättningsverktyg.

En annan förbättring som skulle underlätta för användaren av Turnitin kan påpekas. Då en PDF-fil skickas in till Turnitin borde det framgå i Turnitins gränssnitt om den är skyddad eller inte. Detta för att en skyddad PDF–fil inte kan granskas av Turnitin, men i dagsläget märks inte detta för användaren. Turnitin kommer meddela att texten i filen har en matchningsprocent på noll, vilket nödvändigtvis inte alltid är fallet. Om lärare och andra användare ej har denna vetskap om verktyget, kan detta resultera i felaktiga bedömningar av uppsatser som egentligen är plagiat.

11.3 Felkällor

Då tiden har varit begränsad har det inte funnits möjlighet att göra så många tester som önskats. Det visade sig att tiden för att utföra alla modifikationer i en text hade underskattats från början. Därför har endast ett tiotals texter modifierats och undersökts. Hade fler tester gjorts i varje undersökning hade slutsatserna varit mer tillförlitliga. Då det är få tester, är risken större att slumpen har fått alltför stor inverkan. Detta medför att slutsatserna från resultaten kan vara aningen felaktiga och missvisande.

11.4 Slutsats

Då syftet med denna uppsats var undersöka möjligheten att implementera ett verktyg som kan "lura" plagieringskontrollsverktyg, kan det slutligen konstateras att detta är möjligt. Detta påstående gäller främst Turnitin, men det finns dock ingen garanti för att det med hundra procents säkerhet ej genomskådas av Turnitin.

Litteraturförteckning

[1] Prabhakar Raghavan & Hinrich Scütze Christopher D. Manning. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[2] Ronald A. Cole. Survey of the state of the art in human language technology. `http://www.lt-world.org/hlt-survey/master.pdf`, Mars 2011.

[3] Google. Företagsinformation. `http://www.google.com/corporate/tech.html,`, April 2011.

[4] Viggo Kann. Föreläsning i språkteknologi, statistiska metoder 3, ordklasstaggning, 2010.

[5] Jurafsky & Martin. *Speech and language processing – An Introduction To Natural Language Processing, Computational Linguistics and Speech Recognition*. FIXA, 2009.

[6] Nationalencyclopedin. Plagiat. `http://www.ne.se/plagiat`, Februari 2011.

[7] University of Technology Sydney. Why students plagiarise. `http://www.iml.uts.edu.au/assessment/plagiarism/why.html`, Februari 2011.

[8] Chris Park. In other (people's) words: plagiarism by university students - literature and lessons. `http://www.lancs.ac.uk/staff/gyaccp/caeh_28_5_02lores.pdf`, Februari 2011.

[9] PlagiarismDetect.com. About accurate account. `http://www.plagiarismdetect.com/about-accurate-account.html`, April 2011.

[10] PlagiarismDetect.com. About demo account. `http://www.plagiarismdetect.com/about-free-account.html`, April 2011.

[11] plagiarism.org. `http://www.plagiarism.org`, April 2011.

[12] Plagiarism.org. What is plagiarism. `http://www.plagiarism.org/plag_article_what_is_plagiarism.html`, Februari 2011.

[13] Refero. Vad är plagiering? `http://www.bi.hik.se/Refero/tutorialFlash/2plagiarism.php`, April 2011.

[14] Språkteknologi.se. Vad är språkteknologi? http://sprakteknologi.se/vad-aer-sprakteknologi, Mars 2011.

[15] Svenskuppslagsbok. http://svenskuppslagsbok.se/tag/homonym, April 2011.

[16] turnitin. http://www.turnitin.com, April 2011.

[17] Uppsala Universitet. http://info.uu.se/uadm/dokument.nsf/sidor/7D6CD5D71927EB48C12572D6004D0A91?OpenDocument, Februari 2011.

[18] Wikipedia. Svalor. http://sv.wikipedia.org/wiki/Svalor, April 2011.

[19] Peter Witasp. Mail Konversation.