Dölja det dolda meddelandet

JARI KEMI
och ELIAS THIL

KTH Datavetenskap
och kommunikation

Examensarbete
Stockholm, Sverige 2011

Dölja det dolda meddelandet

J A R I K E M I
o c h E L I A S T H I L

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

URL: www.csc.kth.se/utbildning/kandidatexjobb/datateknik/2011/
kemi_jari_OCH_thil_ elias_K11070.pdf

Kungliga tekniska högskolan
Skolan för datavetenskap och kommunikation

KTH CSC
100 44 Stockholm

URL: www.kth.se/csc

Abstract

The subject of this essay is about a variety of ways to hide information within a message, sent between two specific parties, from any eavesdropper for whom the message was not indented. This hiding of information is called Steganography. There is an introduction regarding the subject involving famous examples of hiding messages where one of them involves the printing of messages under the skin over areas that are normally covered with hair. When this area later is shaved the message is revealed. This is one of the oldest ways of known use of the principles of steganography.

The section with historical examples is followed by more specific information regarding the application of stenganography, mainly considering aspects of the present day and in association with the technology that exists. Special attention is given to the aspects of transferring hidden messages over the Internet, where the messenger can be represented by a picture or just a chunk of normal text. The computers come to play a role in the creation and decryption of these steganograms. A discussions follows involving the increasing power in today's computers and if it might be possible to create a steganogram that can only be discovered by a human and not a computer, explaining what parameters that might be associated with these differences.

Statement of collaboration

This document is an essay written for the examination of the so called 'Degree Project in Computer Science, First level' which is being provided by the Royal Institute Of Technology in Stockholm, Sweden.
The information presented in this essay is provided by the students Jari Kemi and Elias Thil during the spring of 2011 and the compiled facts, the reasoning and the overall information is supported by the sources presented at the end of this document. The initial work behind this essay lied in canvassing the information provided by the sources and connecting related sections of these to the specifics of the purpose for this essay. During this phase of the project, involving this essay, both students were equally involved in finding the facts as well as the distinction of related material.

Following this initial phase of fact finding Elias took it upon himself to start putting the gathered information about steganography together and create a first version of the essay including the introduction and the explanatory parts as well as creating some of the associated pictures for the essay. During the same period of time Jari began to gather the information related to the more specific aspects of lexical steganography and developing our own way of presenting a systematic approach to the subject and concluding the possibilities as well as the limitations that are involved in the process of creating and reading these stegosystems.

Later the final essay was created by a mutual effort from both Jari and Elias where the gathered information and established results and conclusions where brought together and presented in this essay.

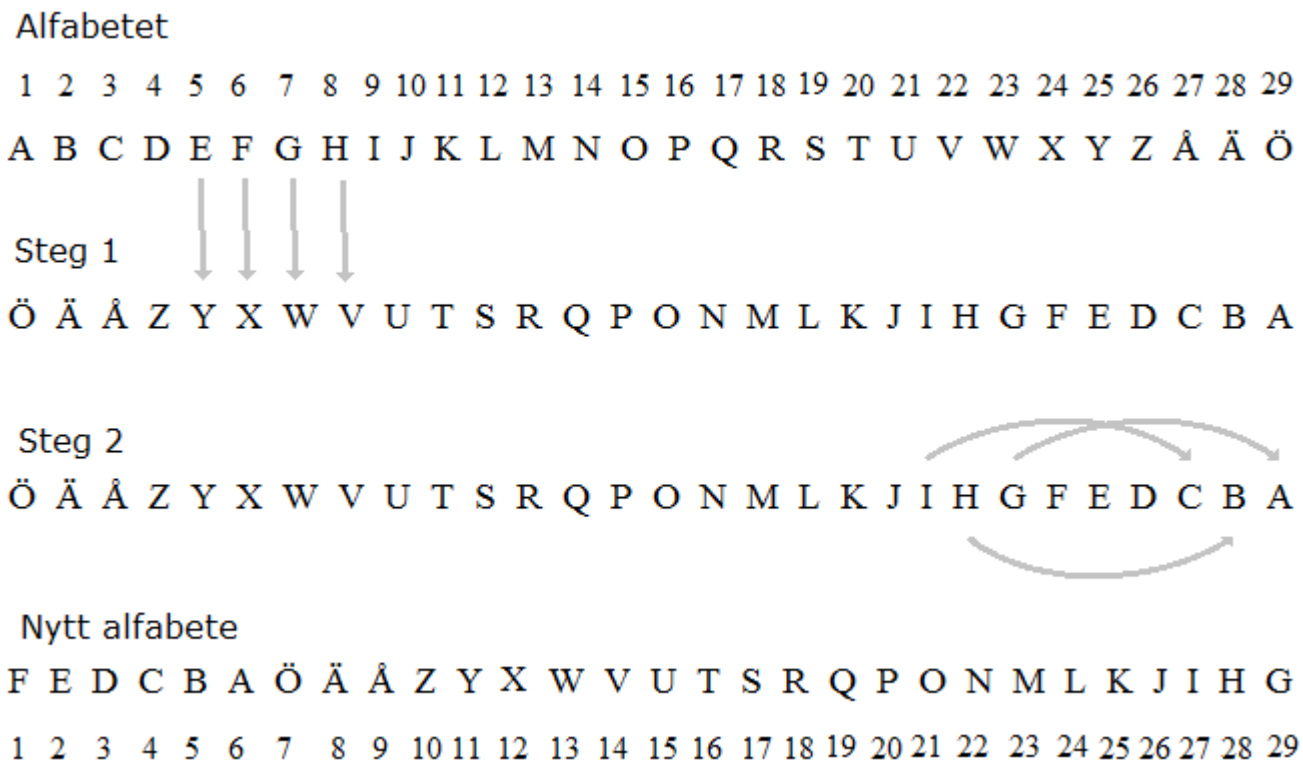
Innehållsförteckning

1.1 Inledning. 6
2.1 Problemställning. 7
3.1 Steganogram. 7
4.1 Historisk bakgrund 9
5.1 Steganografi idag. 10
5.2 Steganografi i ljud.. 10
5.3 Steganografi i bild. 11
6.1 Steganografi i naturlig text. 12
6.2 Antal möjliga kodade bitar i text. 16
6.3 Analys av synonymer 17
6.4 Tvetydighet. 19
6.4.1 Inledning. 19
6.4.2 Tvetydighet i ord. 19
6.4.3 Textjämföring. 29
7.1 Sammanfattning. 22
8.1 Källor. 23

1.1 Inledning

Steganografi härstammat från det grekiska ordet steganos som har betydelsen dold eller skyddad och just detta är vad det handlar om när man applicerar steganografi till existerande meddelanden. Den idag mer allmänt kända varianten för att konstruera hemliga och privata meddelanden är via kryptografi, där ett meddelande skrivet i klartext som ska skickas först genomgår en process av att det läsbara meddelandet omstruktureras med hjälp av en specifik algoritm. Denna omstrukturering resulterar i ett totalt obegripligt meddelande som dock fortfarande innehåller det ursprungliga meddelandet, om man först lyckas nysta upp det medvetet skapade oredan. För att återskapa meddelandet är man oftast tvungen att ha tillgång till en avsatt nyckel kallad kryptonyckel.

I ett illustrerande syfte så antar vi ett scenario där en spion i fiendeland har kommit över information om en omfattande attack mot hemlandet. Denna spion ska nu varna sin handläggare som sedan ska föra meddelandet vidare. Meddelandet han ska föra vidare lyder 'kosacken anfaller från norr' och struktureras om genom att tillverka ett nytt alfabete.



Processen för att skapa detta kryptogram genomgår två steg där alfabetet först spegelvänds varpå det sedan förskjuts sex bokstäver åt höger. Om man numrerar det ordinarie alfabetet och sedan det nya alfabetet så bildar man den kryptonyckel som sedan används för att skapa och lösa upp kryptogram inom just denna ram av konstruktion.

Det nya meddelande som bildas efter att processen är färdig lyder 'zuqfdzbv fvafyybr ariv vurr' och det är givetvis helt omöjligt att utläsa information ur meddelandet som det står i sin nya version. Ett kryptogram likt det vi nu skapat kan dock ibland lösas genom att återskapa nyckeln genom diverse algoritmer och metoder. Ett system likt det ovan är en enkel version som kan knäckas relativt enkelt genom att studera mönster och statistik i språket. Genom detta kan man komma fram till sannolikheten för förekomsten av olika bokstäver i svensk text. Att lösa ut meddelandet med hjälp av statistik blir givetvis enklare ju längre text som finns tillgänglig men det är bara en variant av många sätt att angripa problemet på.

Ett scenario där detta kryptiska meddelande läses av någon annan än spionens avsedda handläggare skulle givetvis kunna resultera i att hans hemliga uppdrag misslyckas och att hans täckmantel är avslöjad. Skulle spionen däremot välja att dölja det krypterade meddelandet 'zuqfdzbv fvafyybr ariv vurr' och skicka det till sin handläggare så skulle det risken för upptäckt kunna minska. Det är detta som innefattas med just steganografi, att dölja det dolda meddelandet.

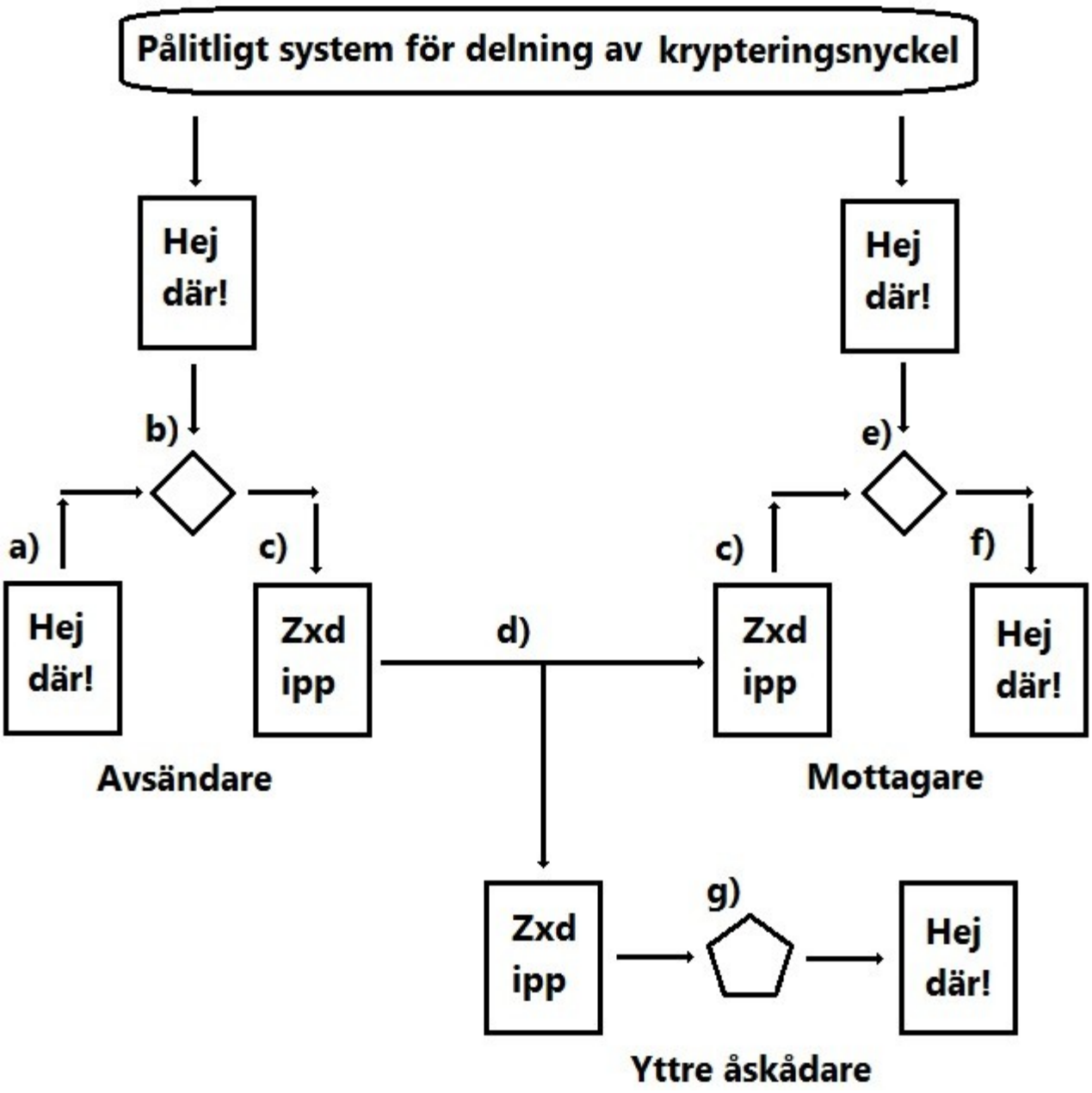
2.1 Problemställning

Syftet med denna uppsats är att identifiera vad som innefattas av begreppet steganografi och undersöka vilka metoder som existerar för att tillämpa steganografi för att dölja dolda meddelanden i vanlig text. Dessa metoder ska sedan analyseras för att nå slutsats kring eventuella begränsningar och möjligheter som kan skilja dem åt. Det ska även tas reda på om det finns ett mer utvecklat och fördelaktigt alternativ för att dölja ett meddelande i naturlig text. Avslutningsvis så ska en utförlig analys av två aspekter inom lexikal steganografi utföras och presenteras.

3.1 Steganogram

Vid konstruktionen av ett stegosystem utgår man alltså från ett första meddelande som ska skickas från en avsändare till en mottagare. Meddelandet krypteras med hjälp av en nyckel och sen döljs. Nyckeln delas antingen av avsändaren och mottagaren eller så har de nycklar som motsvarar mer specifika versioner av kodning och avkodning. Dessa nycklar är ytterst viktiga att hålla hemliga och bör överföras via en säker kanal för trafik, för att sedan kunna möjliggöra ‘säker’ överföring via osäkra kanaler.

Låt oss säga att ett meddelande M ska krypteras med en nyckel K och sedan skapa resultatet R, då kan denna process beskrivas med formeln M*K=R. När mottagaren sen tar emot R behöver hon således tillgång till åtminstone nyckeln K så att R*K=M. Funktionellt kan det då beskrivas som f(M,K) = R och f(R,K) = M. Andra system kan i sin tur kräva att även ett ursprungligt dokument eller bakomliggande material för steganogrammet används för att får fram meddelandet.



a) Det ursprungliga meddelandet i klartext
b) Kryptering av meddelandet
c) Det krypterade och olästa meddelandet
d) Överföring mellan avsändare och mottagare, sårbar för avlyssnyng
e) Avkodning av meddelandet
f) Det ursprungliga meddelandet i klartext
g) Försök till att knäcka det kodade meddelandet

När steganogrammet sedan skickas mellan parter så är det ofta möjligt för utomstående att granska det som överförs och beroende på hur kommunikationen fungerar, tillsammans med val av budbärare för informationen, så kan det existera begränsningar i systemet. Ett stegosystem kan ha passiva åskådare i den mån att de inte har möjlighet att manipulera den information som skickas över kanalen men det finns också scenarion där detta faktiskt är fallet. Säg att två fångar som spenderar sina dagar i celler som är separerade från varandra kommunicerar genom lappar de lämnar i matsalen vid lunch- och middagstid. När fångvaktarna upptäcker deras system med lappar så kanske de inledningsvis inte lyckas lösa ut kryptogrammet men har då möjlighet att rekonstruera meddelandet innan den avsedda mottagaren får möjlighet att avläsa det. Denna typ störande åskådare kan alltså ha stor påverkan på ett steganogram och för att försöka skydda sig mot att informationen går förlorad vid manipulation så står man inför en stor utmaning. Samma typ av störande manipulation av steganografiskt dolda meddelanden kan innefattas av byta filmformat på bilder, översätta textstycken eller komprimera datafiler.

4.1 Historisk bakgrund

En av de äldre historierna om användandet av steganografi härstammar från antikens Grekland där meddelanden skickades mellan parter nedskrivna på vaxtavlor. Dessa vaxtavlor bestod ursprungligen av en träskiva täckt i vax som sedan kunder återanvändas som skrivyta. Hemliga meddelanden kunde skickas genom att man ristade in ett meddelande på själva träskivan innan man dränkte den i vax och sedan skrev ett helt oskyldigt meddelande över dess yta. När vaxtavlan nådde sin avsedda mottagare kunde denna sedan avtäcka det dolda meddelandet utan att någon som eventuellt granskat den under transport anat dess egentliga budskap.

Historier om tyrannen Histiaios som levde under 500-talet f.Kr. berättar om hur denne efter att ha lyckats lura fienden med att han övergav sin tidigare kung och anslöt sig till fienden. Han kom senare att tatuera in ett meddelande på skalpen av sin mest pålitliga slav och lät sedan håret på denne växa ut igen. Efter en tid som rådgivare och vän för fienden skickades slaven sedan iväg varpå mottagaren uppmanades att raka slavens huvud som sedan framförde budskap om att göra uppror mot perserna.

Generellt har det ur historisk synpunkt funnits ett speciellt intresse för bruket av steganografi under krigstider varpå man under andra världskriget bland annat använde sig av att markera bokstäver i text som tillsammans byggde upp ett kryptogram. Dessa maskinskrivna bokstäver var inledningsvis markerade genom att vara aningen förskjutna i höjdledd på pappret men kom senare även att markeras med små punkter, antingen skapade med pyttesmå hål eller markeringar med osynligt bläck.

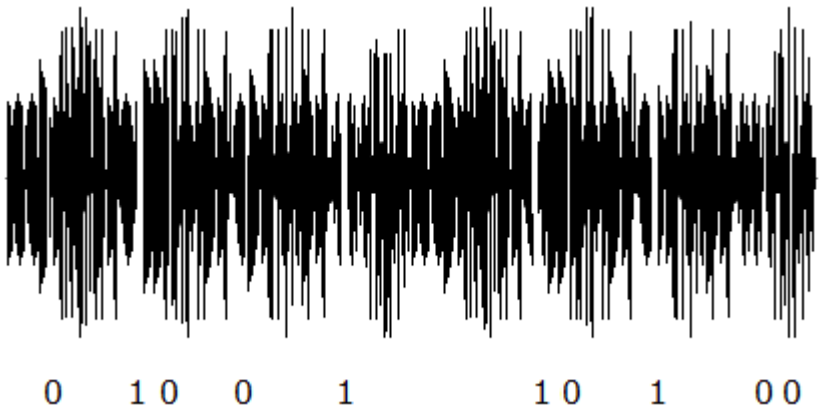
5.1 Steganografi idag

Nu för tiden finns det många sätt att tillämpa steganografi och likaså de syften och sammanhang vari det kan komma till stor användning. I industriella sammanhang finns det idag dolda meddelanden inuti pengasedlar i form av invävda magnetremsor och det finns vattenstämplar i copyright-skyddat material som distribueras inom musik- såväl som filmindustrin. Denna typ av vattenstämplar kan konstrueras på många olika sätt och har ofta för avsikt att fungera i ett identifierande syfte när det gäller att finna källan för eventuellt internt såväl som externt läckage under tillverkningsprocessen. Detta gäller exempelvis när filmer och musik som distribueras internt senare återfinns på internet och via ett system av olika vattenmärken kan man avläsa vilken av de kopior man delat ut internt det är som återfinns på internet.

5.2 Steganografi i ljud

När man ska konstruera ett steganogram inuti en ljudfil kan man exempelvis konstruera ett system av sk. ekon inuti själva ljudupptagningen. Tanken är att man skapar väldigt små tomrum i flödet som förslagsvis är av längden 0,5 ms och 2 ms som sedan bygger upp en följd av tomrum. Dessa ekon kan enkelt konstrueras med en stor mängd gratis digital mjukvara som idag finns tillgänglig över internet. Denna typ av manipulering görs också utan att visa någon extern förändring i vare sig storlek på själva filen eller i det hörbara resultatet från en uppspelning av den. Det mänskliga örat klara inte av att urskilja avbrott kortare än 3ms. Om konstruktionen av steganogrammet lyckas så kan ekona inuti ljudfilen dock enkelt avlösas av motsvarande mjukvara på datorn igen. Enklast vore då att låta de olika avbrotten motsvara digitala ettor och nollor som sedan kan bygga upp korta såväl som längre meddelanden.

Inledningsvis tar man alltså en redan existerande ljudupptagning, förslagsvis en musikfil redan existerande på datorn, och manipulerar in dessa typer av tystnad in i sekvensen av ljud. För själva avläsningen av steganogrammet använder man enklast den ursprungliga ljudfilen som mall och noterar antingen grafiskt eller automatiskt via specifik ansluten mjukvara de skillnader som existerar inuti uppspelningen. Här följer en illustration av en orörd ljudsekvens följd av en manipulerad sådan där den tio bitar långa sekvensen 0100110100 är inbyggd vari avbrott på 0,5 ms motsvarar en binär 0:a och de på 2 ms motsvarar binära 1:or.



Kommunikation mellan fångar samt deras kontakt med yttre världen har kommit att utveckla flera aspekter inom steganografi av naturlig text med. Ett äldre exempel är hur fångar skickade ut meddelanden i form av morsekod via modifiering av punkter och apostrofer hos bokstäverna i, j t och f. Dessa variationer existerade i vanliga handskrivna brev som skickades ut och in från fängelse. I dagens läge bl.a. inuti amerikanska fängelse så sitter gängledare och gangsterbossar fängslade utan att bli av med sin höga position inom den undre världen. Dessa fångar utvecklar allt mer komplicerade metoder för att skicka ut meddelanden och uppmaningar till sina kriminella undersåtar och fortsätter således vara aktivt kriminella även inifrån fängelset.

En annan variant av steganografi som ligger i tiden är hur militär och volontärer i krigsdrabbade zoner tränas i signalement med kroppsspråk samt text för att kunna bilda ord och meddelanden under gisslansituationer och förmedla extra information till sina mottagare. Ett exempel är hur en gisslan på film med skakiga händer kan fippla på bokstäver på sin t-shirt, markera kroppsdelar med lätta knackningar på sin kropp och diverse ansiktsuttryck medan de pratar in meddelanden på de filmer som sedan används vid utpressning. I handskrivna brev finns det också stor möjlighet för att manipulera den information man skickar vidare exempelvis med specifikt ordval och variationer av synonymer eller likt systemet med manipulation av specifika bokstäver så kanske man lyckas sända iväg koordinater samt viktiga uppgifter utan att ens gisslantagare uppmärksammar det.

5.3 Steganografi i bild

De bästa steganografiska systemen idag använder sig av grafiska verktyg för att modifiera existerande bilder för att dölja krypterad data. Den enklaste metoden är nog genom att koda in en bit från kryptogrammet i utvalda pixlar genom att modifiera vardera pixels så kallade RGB-värde, som bygger upp färgdjupet av röd, grön och blå i just den pixeln. Dessa värden representerar var och en av en byte data med värden mellan 0 och 256 som

motsvarar olika styrka. Genom att ändra byte-värdets minst signifikanta bit med kryptogrammets värde, så påverkar man pixelns visuella presentation ytterst lite och detta är något som ofta inte kan uppfattas av det mänskliga ögat. Detta resulterar i ett smidigt och väldigt effektivt bruk av steganografi.

6.1 Steganografi i naturlig text

Lexikal steganografi är ett område och en vetenskap som ännu är väldigt mycket i sin ungdom och det finns inga etablerade eller standardiserade metoder för att utföra den process som krävs för att bädda in ett meddelande i en redan färdig text. Av de metoder som idag existerar på en mera experimentell nivå så är det av stor vikt att bevara innebörden av den ursprungliga och ev. modifierade texten.

Vårt arbete är specificerat kring användningen av synonymer och hur man genom att manipulera meningar i en existerande text kan tillsammans med synonymer för specifika ord dölja ett bakomliggande budskap.
Många ord har mer än en innebörd och det är sällsynt för två ord att vara synonymt i alla sina aspekter. Två synonymer kommer sannolikt att ha mycket olik syntaktisk distribution och frekvens i en viss text eller i ett visst sammanhang. Detta leder till svårigheten att skapa steganogram med flytande och trovärdig text.

En typ av system som skulle kunna användas för att effektivt dölja meddelanden i vanlig naturlig text utan att dra till sig uppmärksamhet kan realiseras genom strategiskt användande av synonymer. Ett sådant steganogram kan skapas med syfte att vardera ord som är utbytbart har en bakomliggande betydelse i kontexten. Eftersom språket i huvudsak är mångtydigt, kommer det att finnas flera sätt att skriva en mening på. Med hjälp av steganografi kan vi bädda in data i naturlig text och kunna bevara textens betydelse, beroende på vilken strategi man använder. Ha i åtanke att det språk ofta kan alterneras och betrakta följande mängd M av meningar:

*M = { Arvid **hoppade** över staketet för att **leka** med sina vänner,*
*Arvid **hoppade** över staketet för att **busa** med sina vänner,*
*Arvid **skuttade** över staketet för att **spela** med sina vänner,*
*Arvid **skuttade** över staketet för att **lira** med sina vänner }*

Om de möjliga synonymgrupperna S_i i föregående mängd av meningar är S_1={hoppade, skuttade} och
S_2 = {leka, busa, spela, lira}, så kan man använda synonymiteten för att generera sammanlagt åtta meningar. Detta eftersom vi har två stycken synonymgrupper att välja ifrån där den första gruppen innehåller två ord, **hoppade** och **skuttade,** och den andra, **leka**, **busa**, **spela** och **lira,** innehållande fyra ord.

Vi kan skriva meningarna som:

*Arvid **hoppade** över staketet för att **lira** med sina vänner.*
*Arvid **hoppade** över staketet för att **leka** med sina vänner.*
*Arvid **hoppade** över staketet för att **busa** med sina vänner.*
*Arvid **hoppade** över staketet för att **spela** med sina vänner.*
*Arvid **skuttade** över staketet för att **lira** med sina vänner.*
*Arvid **skuttade** över staketet för att **leka** med sina vänner.*
*Arvid **skuttade** över staketet för att **busa** med sina vänner.*
*Arvid **skuttade** över staketet för att **spela** med sina vänner.*

Genom att tilldela data för specifika ordval så skulle ett effektivt sätt att gå till väga på vara att ge vardera ord i en synonymgrupp ett binärt värde. För att sedan ha möjlighet att konstruera alla möjliga följder av binära strängar så måste man ha detta i åtanke när man skapar sitt 'binära alfabete'. En metod vore att ge alla orden i en mängd av n ord ett värde mellan 0 och n-1. Således får vi från de fyra synonymerna binära talen 00, 01, 10 och 11 som då är individuellt bundna till ett specifikt synonym. För meningen "Arvid **hoppade** över staketet för att **leka** med sina vänner", så fördelas bitarna med denna metod på följande sätt:

*Arvid **hoppade 0** över staketet för att*	***leka***	***00***	*med sina vänner*
skuttade 1	***busa***	***01***	
	spela	***10***	
	lira	***11***	

Med denna mening kan vi koda in en tre bitar lång binärsträng genom att slå ihop båda bitsträngarna. Om vi skulle vilja koda in bitsträngen '101' in i meningen, representerat som talet 5 i decimala siffror, så väljer man ordet med bitsträngen '1' från första valet och från det andra synonymgruppen väljer man ordet med bitsträngen '01'. Detta resulterar i meningen 'Arvid skuttade över staketet för att busa med sina vänner'.

Med denna metod är det inte alltid möjligt att ge varje ord i en synonymgrupp en unik bitsträng eftersom man måste kunna koda in alla de möjliga sekvenserna. För att varje ord i en synonymgrupp ska kunna ha ett unikt n-siffrigt binärsträng så måste synonymgruppen innehålla minst 2^n ord eftersom ett binärt tal med n bitar kan anta 2^n möjliga n-siffriga binära tal.

Betrakta följande mening:

*Johanna **hittade** en leksak*

Låt oss säga att man fann en synonymgrupp S som är ansluten till ovanstående mening där S = {hittade, fann, lokaliserade, påträffade, upptäckte}. Nu har vi istället tillgång till fem synonymer och med fler ord så följer möjligheten att tilldela ett större antal siffror. Om man dock eftersträvar att skapa ett funktionellt system för bruket av synonymer och tilldelade värden inom synonymgrupperna så stöter man på begränsningar med det binära talsystemet. För att kunna vara konsekvent med längden på meddelanden och delar av den så måste alla binära delsekvenser ha samma antal bitar. För att i vårt fall kunna möjliggöra ett femte tal så måste vi alltså lägga till en binär siffra på samtliga tal för att uppnå ett enhetligt system för tilldelade värden till orden.

Johanna	***hittade***	***00***	*en leksak*
	fann	***01***	
	lokaliserade	***10***	
	påträffade	***11***	
	upptäckte	***??***	

I detta fall, så har vi möjligheten att ge ordert 'upptäckte' vilket värde som helst eftersom vi har redan uppfyllt kravet att kunna koda alla möjliga tvåsiffriga binärvärden. Om ordet ges ett värde med en bit mer t.ex. värdet '100' så skulle man möjligtvis kunna lagra mer data, men sannolikheten att just det ordet kodas är mindre än för de orden med två bitar. Säkerheten i texten kan stärkas genom att tilldela ordet 'upptäckte' ett redan existerande binärt värde vilket leder till möjlighet till skaparen av steganogramet att välja mellan ord som har samma värde.

Johanna	***hittade***	***00***	*en leksak*
	fann	***01***	
	lokaliserade	***10***	
	påträffade	***11***	
	upptäckte	***11***	

Med denna begränsning för tillgång till binära värden så kan man alltså fortfarande utnyttja de överflödiga orden till att främja syntaxen och den kontext vari synonymerna förekommer. Författaren för steganogrammet får

således en större möjlighet att utnyttja den tillgängliga synonymgruppen och dessa dubbletter av bitvärden skulle givetvis kunna utnyttjas vidare för att få en text att se mindre misstänksam ut. Betrakta nu följande mening:

*Det var ett **enkelt** problem*

Låt oss säga att man från databasen av synonymer fann en synonymgrupp S som är ansluten till ovanstående mening där S = {enkelt, lätt, simpelt, elementärt, primitivt, spartanskt, torftigt}. Synonymgruppen S1 innehåller sju ord, detta är ett ord för lite för att man ska kunna ge varje ord en tre bitar lång binärsträng. Synonymgruppen skulle således kunna konstrueras på följande sätt:

S = {	***(enkelt, lätt)***	***00,***
	(simpelt, elementärt)	***01,***
	(primitivt, spartanskt)	***10,***
	(torftigt)	***11}***

Ur systemet ovan så kan man avläsa att orden för bitsrängarna ‘00’, ’01’ och ‘10’ kan väljas på två sätt, medan ordet för bitsträngen ‘11’ endast kan väljas på ett sätt, nämligen ordet ‘torftigt’. Nu kan generatorn för steganogrammet ge författaren en valmöjlighet av ord för dessa tre binärsträngar. Låt oss säga att man önskar koda in bitsträngen ‘01’ i meningen ovan. Detta skulle alltså resultera i två möjliga fall:

*Det var ett (**simpelt**, **elementärt**) problem*

Avslutningsvis kan det vara klokt att jämföra innebörden av den ursprungliga meningen ‘*Det var ett enkelt problem’* och med meningarna ‘*Det var ett **simpelt** problem’* och ‘*Det var ett **elementärt** problem’*. Man inser kanske att ordet ‘simpelt’ representerar innebörden av ursprungliga meningen bäst och i steganogrammet kommer meningen slutligen att bli:

*Det var ett **simpelt** problem*

När det gäller just denna typ av steganografi där man har ett bredare val av antal synonymer så används det ofta ord som förekommer frekvent i vanliga texter. Som vi sett så ökar tröskeln för hur många binära siffror vi kan inkludera i en synonymgrupp med faktor två för antalet synonymer och det finns uppenbara praktiska begränsningar för hur många binära siffror man kan bygga in i ett steganogram genom att utöka vardera

synonyms värde. Istället så vore således en begränsning på två siffror säkerligen ett positivt inslag till ett steganogram då det lämnat utrymme för skaparens inflytande och valmöjlighet att anpassa sig till kontexten för att enklare skapa ett säkert steganogram och det är ju trots allt detta som är det främsta syftet med steganografi.

6.2 Antal möjliga kodade bitar i text

Ett ersättningssystem för ord bör ha möjlighet att variera krypteringens densitet över den text som behandlas. Genom att glest fördela de specifika orden för krypteringen på en stor text så att manipuleringar inte är koncentrerade i ett litet område så får man som resultaten låg krypteringsdensitet.
Den sammanställda texten kommer därför att vara mindre uppseendeväckande för en mänsklig läsare.
Ett system med större täthet av kryptering resulterar alltså i att texten löper större risk av att bli såväl grammatiskt inkorrekt och sakna sin ursprungliga innebörd.

Om en text innehåller n st utbytbara ord betecknade s1, s2, ..., sn vars synonymgrupper kan anta S1,S2,...,Sn antal ord, så är dess totala kapacitet av kodbara bitar:

$$C = \log_2\left(\prod_{i=1}^{n} \left| S_i \right|\right)$$

Om antalet dolda bitar att kodas är b, så kan man dölja meddelandet C / b gånger, vilket kallas modulationsfrekvens. Låt oss anta att en text innehåller åtta möjliga synonymgrupper att lagra data med. Säg att dessa synonymgrupper innehåller (4,4,2,8,8,4,8,4) möjliga val. Detta leder till:

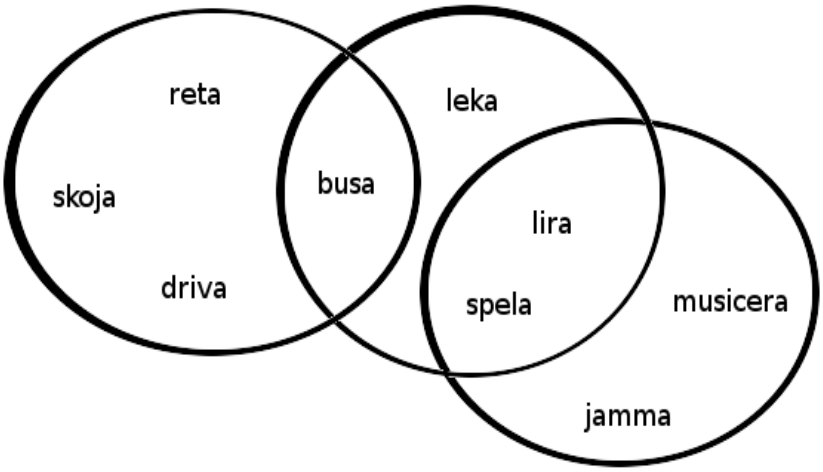
$$C = \log_2(4*4*2*8*8*4*8*4) = 18$$

Låt oss säga att vi vill koda in en bitsträng med nio bitar tex ‘101101100’, då blir modulationsfrekvensen 18/9 =2. Detta leder till att vi i genomsnitt måste koda ett ord per två av de kodbara orden.

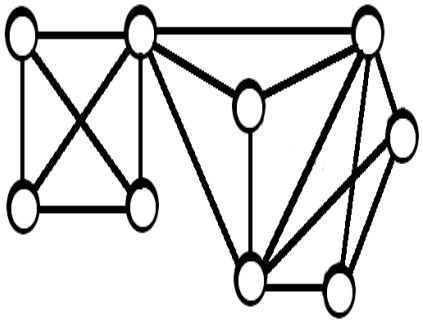
6.3 Analys av synonymer

När det kommer till att använda ett system av synonymer för att konstruera ett steganogram i en existerande naturlig text kan man utföra det på flera olika sätt. Inledningsvis kan man använda sig av de tillhörande synonymerna och utgå från en existerande text för att sedan skapa en ny text som innehåller utbytta ord. En mottagare som har tillgång till originalet kan enkelt finna de ord som ändrats och sedan avkoda det meddelande som dessa ord motsvarar.

Ett inledande problem med just bruket av synonymer är hur det kan förekomma koalition mellan hopar av synonymer som sins emellan delar synonymer. Något som gäller för de flesta språk för användandet av synonymer är förekomsten av överlappningar inom synonymgrupper, vilket kan leda till omfattande problem vid bruk inom steganografi. Dessa framkommer när vi läser in ett steganogram och identifierar ett ord som vi vet med oss är knutet till ett värde. Låt oss säga att ordet tillhör flera synonymgrupper, hur kan vi veta vad ordet har för kodat värde?



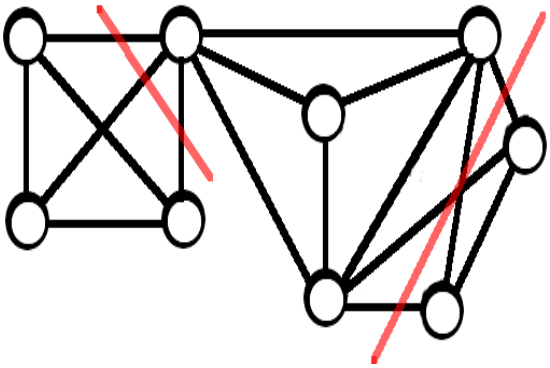
Graf 1: Ord tillhör flera synonymgrupper



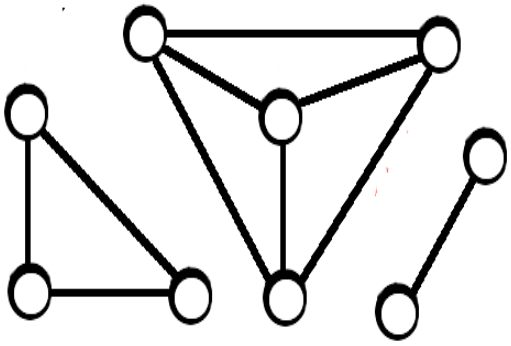
Graf 2: Synonymgrupperna från graf 1 representerad med kanter och noder.

Om ett ord tillhör flera synonymgrupper så är det inte möjligt för avkodaren att veta vilken synonymgrupp ordet i detta fall kom ifrån. Detta leder till att det enda sättet att avkoda meddelandet är genom att med hänsyn till alla synonymgrupper, börja generera alla de möjliga lösningar som kan förekomma. Av alla dessa genererade möjliga meddelanden så letar man sedan upp ett meddelande som har vettig information. En text med upprepade scenarion av överlappningar genererar en exponentiell mängd möjliga meddelanden. Låt oss säga att varje ord tillhör exakt två synonymgrupper. Om vi ändrar på 100 ord i texten, så genererar avkodaren 2^{100} möjliga meddelanden, vilket är ett enormt stort tal. För att använda oss av vår metod måste vi då dela upp

synonymgrupperna så att ett ord tillhör endast en synonymgrupp. Med långa meddelanden är alltså denna metod ej applicerbar i verkligheten eftersom antalet möjliga meddelanden blir exponentiellt stort.



Graf 3 :Synonymgrupperna som graf med markeringar var de kan beskäras för att kunna skapa kompletta synonymgrupper.



Graf 4: Alla ord tillhör endast en synonymgrupp

Graf 5: Resultatet

Nu har vi blivit av med problemet att identifiera vilka värden som har kodats in i en mening, men vi har också blivit av med flertal möjliga synonymer. Utgå från synonymdatan från de ovanstående bilderna och betrakta följande mening:

*Stefan kan **spela** med gitarren*

Från synonymgrupperna av bilderna ovan hittas en synonymgrupp S från ovanstående mening där S = {spela lira, leka, busa}. Vi har kapat bort orden ‘musicera’ och ‘jamma’ från spelar, vilka är (i detta fall) kontextuellt

bra synonymer för meningen. Nu återstår endast ‘lira’, ‘leka’ och ‘busa’ kvar som synonymer till spelar. Man inser snabbt att om man byter ut ‘spela’ mot ordet ‘leka’ eller ordet ‘busa’ så kommer kontexten av meningen inte nödvändigtvis vara densamma. Följande avsnitt behandlar detta mer ingående.

6.4 Tvetydighet

6.4.1 Inledning

Problemet med tvetydighet är när det gäller lexikal steganografi är av stor relevans. Detta kan spåras tillbaka till frågan ‘Vad är meningen med ett ord?’. Att stegosystem härmar egenheter av naturligt språk kan vara mycket säkerhetsrelevant och som resultat av detta står skaparen av ett steganogram inför utmaningen att bestämma lämplig ersättning av ord för att upprätthålla korrekt meningsuppbyggnad och kontext. Det som resulterar i att ett lexikalt steganogram ska vara säkert så kräver ett omfattande arbete med lexikal språkbehandling och i detta avsnitt kommer vi se varför.

6.4.2 Tvetydighet i ord

I det flesta västeuropeiska språken så finns det väldigt få ‘äkta’ synonymer. Med dessa menas ord som det alltid går att substituera mot varandra i en mening utan att den tillhörande meningens innebörd förändras. Många ord har mer än en innebörd och det är sällsynt för två ord att vara synonymt i alla sina aspekter. Synonymer kommer sannolikt att ha mycket olik syntaktisk distribution och frekvens i en viss text eller i ett visst sammanhang. Detta leder till svårigheten att skapa steganogram med flytande och trovärdig text. Med detta i åtanke, betrakta följande mening med synonymgrupp S för ordet ‘*spelar’*:

Anders spelar poker

S = { lirar, leker, busar }

Kan vi använda varje synonym för ordet ‘spelar’ i denna mening utan att orsaka misstanke om existens av möjliga steganogram i texten? Om meningen ‘*Stefan busar poker’* skulle genereras, så är det inte helt klar vad meningens betydelse är. Med största sannolikhet så skulle kontexten i meningen inte längre vara bevarad. Detta problem är den största nackdelen med att använda ordersättning av synonym som steganografisk modell. Det existerar metoder för detta, men ingen av dessa har hittills lyckats nå en tillräckligt sofistikerad nivå av säkerhet för att garantera att steganogrammet inte ska kunna upptäckas med avseende på brister i kontexten eller meningsuppbyggnaden.

6.4.3 Textjämföring

När det kommer till att byta ut ett ord mot någon av dess synonym så har vi redan etablerat att det kan krävas en delikat process för att systematiskt kunna utföra det utan att utsätta steganogrammet för risker. Dessa risker kan om substitutionen sker vårdslöst leda till enorma konsekvenser och det är i skaparen av steganogrammets största intresse att se till meningsuppbyggnaden kring det utvalda order samt sammanhängande kontext i dess omgivning. I denna uppsats så behandlas möjligheter att jämföra meningar inuti texter och få tillgång till en mer överblickande bild av ett ords förekomst och sammanhang.

Om man vid konstruktion har tillgång till en databas med en stor mängd vanlig text med ursprung från separata källor bör man alltså utföra en sökning efter matchande meningar för synonymerna ur en gemensam synonymgrupp med avseende för deras omgivning i respektive texter. När man hittat ett ord där man har möjlighet att lagra data likt ordet *spelar* i meningen '*Stefan **spelar** med gitarren*' så kan man nu leta efter liknande meningar med dess synonym från databasen med text. Synonymerna i detta fall för 'spelar' skulle kunna vara 'lirar', 'leker', och 'busar'. Nu sökes sedan meningar från databasen som matchar meningarna:

'*Stefan **spelar** med gitarren*'
'*Stefan **lirar** med gitarren*'
'*Stefan l**eker** med gitarren*'
'*Stefan **busar** med gitarren*"

Ut databasen erhölls sedan texterna:

'*Stefan spelar med gitarren*'	'*Stefan lirar med gitarren*'
*...enorm tillgång för en gitarrist som kanske sitter i kontrollrummet och **spelar med gitarren** kopplad direkt till mixerbord eller...* *...problemet är dock att när jag **spelar med gitarren** inkopplad till förstärkaren så händer det...* *...byta strängar beror på hur mycket du **spelar med gitarren** och hur brydd du är om att den ska låta..*	*...min förstärkare eller elgitarr. När jag är i rep-loken och **lirar med gitarren** och sen försöker ta i...* *...känslan och knapplayouten man **lirar med gitarren**. trummorna och micen ska jag inte uttala...* *...sitter på en stol och **lirar med gitarren** i knät. Den blinde sångaren gör en helt okay version av...*

Nu vet man att ‘lirar’ är en lämplig synonym att använda sig av. Däremot hittade vi inte meningar som matchade orden ‘leker’ och ‘busar’ så vi kan endast koda in en bit eftersom vi har en begränsning i form av att vi har endast två ord att välja från. Observera att man även måste hitta en mening som matchar det ursprungliga ordet i meningen som nu i detta fall är ‘spelar’. När avkodaren läser ett möjligt kodat ord så måste den finna alla möjliga ord som kan användas i den specifika meningen. Om avkodaren inte hittar en eller flera matchande meningar av det ursprungliga ordet så finns det en överhängande risk att fel binärsträng kommer att avkodas.

Denna metod medför alltså inte några garantier för att kontexten i det behandlade textstycket bevaras men med medvetet användande så ökar sannolikheten för ett välgjort steganogram. Desto större databas med korrekt text som finns tillgänglig desto större sannolikhet finns det att finna möjliga ersättningar för ord. Det krävs dock att det inte ändras på ord som redan tidigare har matchats med meningar som existerar i samma text då avkodaren löper risk att inte längre kunna matcha orden med varandra för att kunna nå korrekt utdata. Denna metod är relativt enkel att implementera och med vidare utveckling i framtiden så finns det kanske en möjlighet för att kunna konstruera kompletta och säkra steganogram med korrekt språk där sannolikheten för upptäckt av det dolda meddelandet är acceptabelt låg.

7.1 Sammanfattning

När det handlar om steganografi som vetenskap så är det idag ännu ett väldigt outforskat område. Däremot så är användningsområdet för steganografi och varianterna av dess bruk en konstform som länge existerat i den moderna människans samhälle. Användandet för att hemlighålla meddelanden med hjälp av steganografi har ökat stadigt i frekvens i takt med att det moderna informationssamhället utvecklats och i takt med det har i sin tur metoderna för steganografi utvecklats.

När media och information sedan kommit i kontakt med den relativt unga digitala världen så har bruket av steganografi i all dess områden ytterligare mångdubblas och det är idag ett område som studeras mer och mer. Förutom det industriella användningen av steganografi för att märka digital media i form av så kallade vattenstämplar med syfte att motverka piratkopiering så ägnas mycket tid åt analysen av lexikal steganografi.

Då det kommer till att dölja krypterade meddelanden i text så är möjligheterna för att konstruera digitala sekvenser med hjälp av specifika val av synonymer något som är överskådligt och tillämpbart utan att nödvändigtvis dra åt sig uppmärksamhet. När det kommer till uppbyggnaden av meningar och urvalet av synonymer så stöter man dock på begränsningar såväl som möjligheter när det kommer till att bygga upp ett steganogram. Skulle man jämföra lexikal steganografi med andra mera direkta och äldre metoder av steganografi så är det betydligt mer omfattande och invecklat. Om skapandet av ett steganogram dock fullföljs inuti en naturlig text så kommer det ha möjlighet att uppnå en väldigt hög nivå av säkerhet.

8.1 Källor

Towards linguistic steganography

Richard Bergmair, November 10, 2004

A systematic investigation of approaches, systems and issues.

http://richard.bergmair.eu/pub/towlingsteg-rep-inoff-b5.pdf

Informaton Hiding: A survey

Peticolas, Anderson, Kuhn, Juli 1999

http://www.petitcolas.net/fabien/publications/ieee99-infohiding.pdf

Disappearing Cryptography Information Hiding Steganography and Watermarking, 3rd Edition

Peter Wayner, December 2008

Provably Secure Steganography

Hopper, Langford, von Ahn, September 2002

http://www.cs.cmu.edu/~biglou/PSS.pdf