

A project completed as part of the requirements for the
BSc (Hons) of Science of Computing entitled

Towards Linguistic Steganography: A Systematic Investigation of Approaches, Systems, and Issues

by Richard Bergmair

Why Linguistic **Steganography**?

- Cryptosystems can protect sensitive data from unauthorized access, by using a representation that makes a cryptogram impossible to interpret **but**
- they do not conceal the very fact, that a cryptogram has been exchanged

Why Linguistic Steganography?

- this is not a problem, as long as cryptography is perceived at a broad (legal?) basis as a legitimate way of protecting one's privacy, but
- it is a problem, if it seen as a tool useful primarily to potential terrorists.

In order to protect the individual's freedom of opinion and expression, we will have to deal with "Wendy the warden" trying to detect and penalize unwanted communication.

Why Linguistic Steganography?

- Stegosystems can protect sensitive data from being detected, by using a representation that makes steganograms appear as covers (a holiday image, a newspaper article, ...)
- The more covers an arbitrator needs to analyze, trying to detect a steganogram, the more difficult it will get.

Why **Linguistic** Steganography?

- The vast masses of data coded in natural language make for a good haystack to hide a needle in. Steganalytic efforts concentrating on digital images exchanged over the web might still be tractable, but it will hardly be possible to arbitrate all communication that takes place in natural language.
- Natural language messages can easily be transmitted over almost any medium.

Steganographic Security

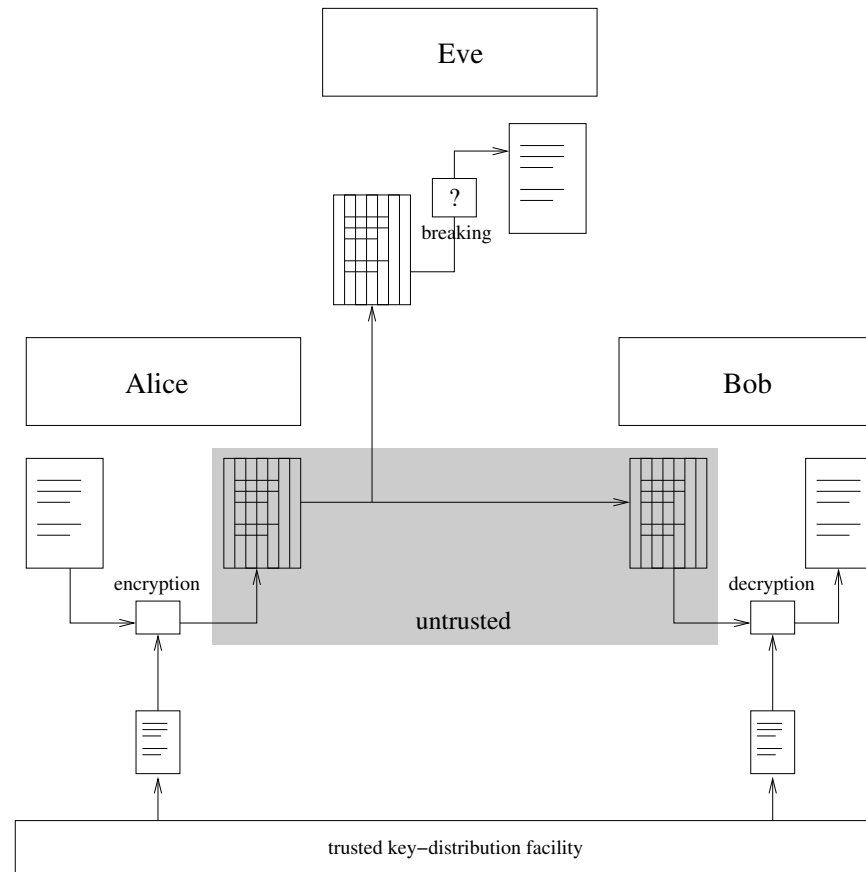
- Alice and Bob want to exchange messages m chosen from a message-space M over an insecure channel. They assume that data submitted over this channel is intercepted by Eve.
- Alice and Bob have a key-distribution facility, which equips them with keys k , chosen from a key-space K . They can safely assume this channel to be secure, in the sense of trusting it, not to expose the keys to Eve.
- Alice and Bob want to make the insecure channel secure, by making the security of the messages depend on the security of the keys.

Steganographic Security

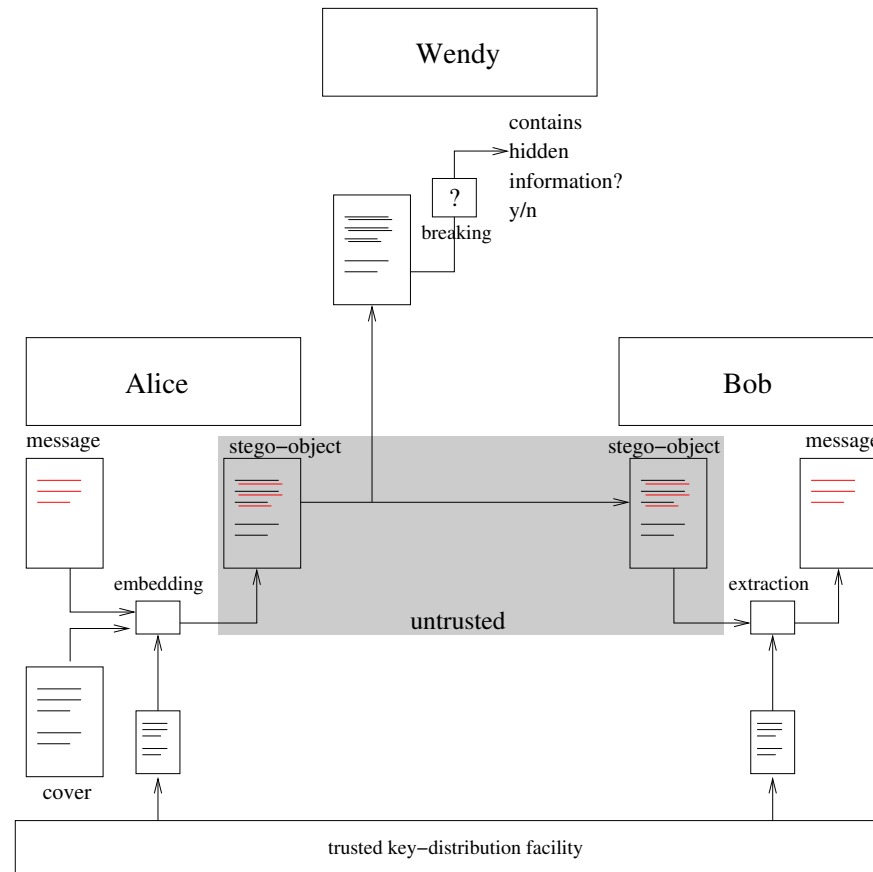
In the cryptographic setting,

- Alice **encrypts** the message m , by choosing a cryptogram e in accordance with the key k :
 $E(m, k) = e$.
- Bob **decrypts** the cryptogram e , i.e. reconstructs the message m from e using k : $D(e, k) = m$. This is possible because $\forall m, k : D(E(m, k), k) = m$.
- Eve tries to **break** the cryptogram. This is impossible because it involves solving a difficult problem.

Steganographic Security



Steganographic Security



Steganographic Security

In the steganographic setting,

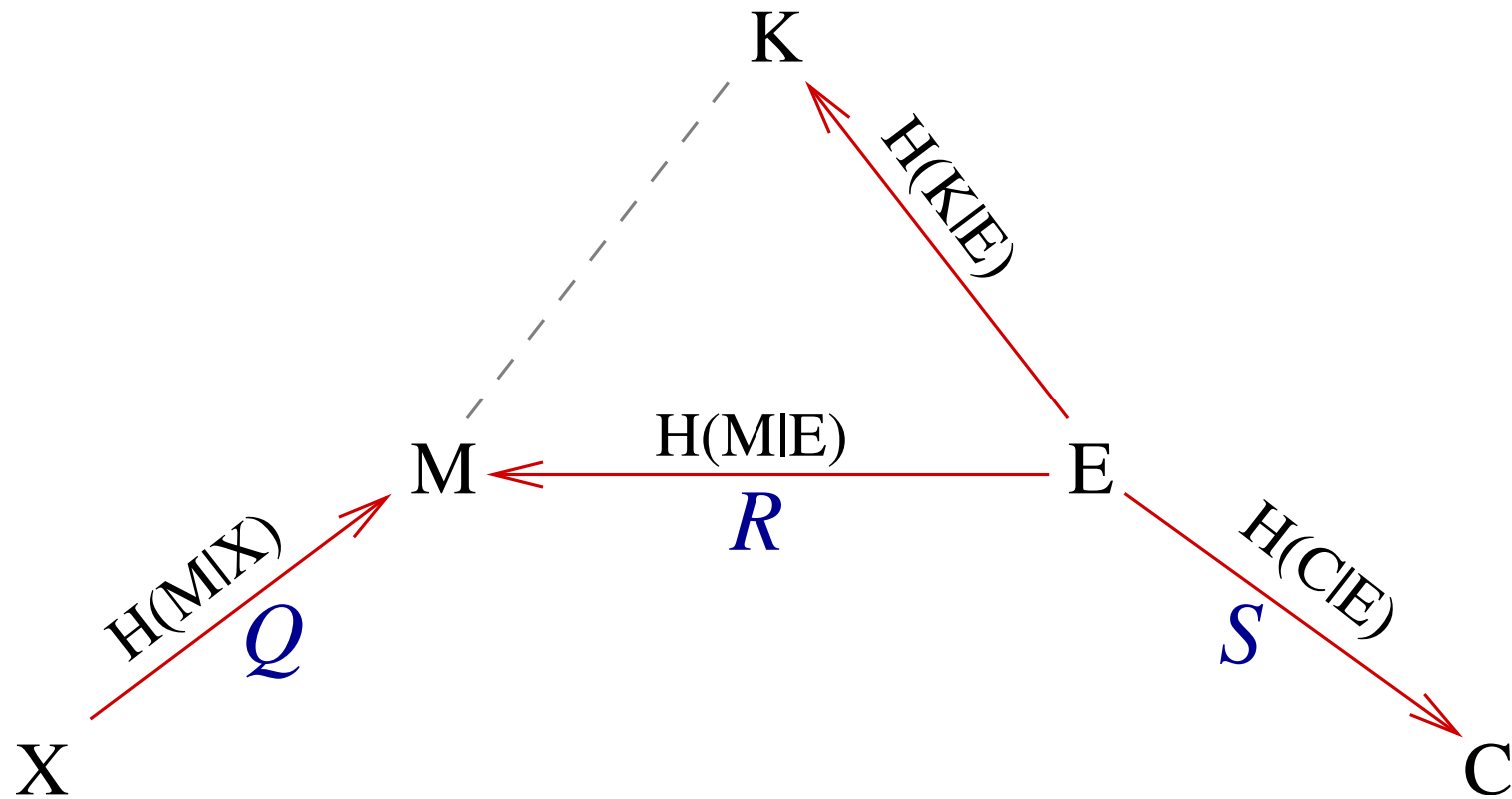
- Alice **embeds** the message m into a cover c , by choosing a steganogram e in accordance with the key k : $E(c, m, k) = e$.
- Bob **extracts** the message from the steganogram e using k : $D(e, k) = m$. This is possible because $\forall m, k : D(E(m, k), k) = m$.
- Eve tries to **detect** the steganogram. This is impossible because there is a cover c' such that the difference between e and c' is **imperceptible** by humans, and machines trying to detect it face a difficult problem in the cryptographic sense.

Steganographic Security

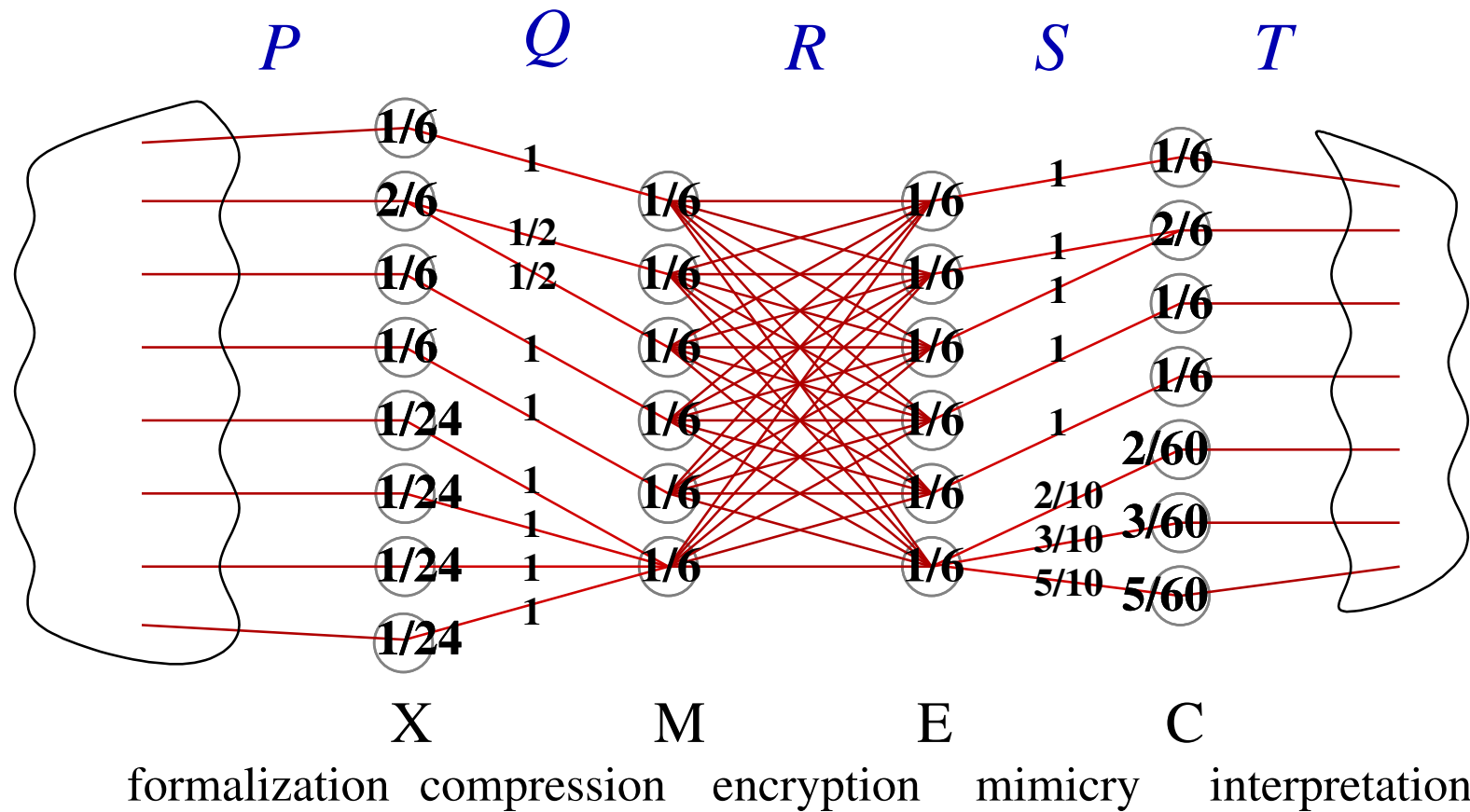
A difficult problem in the cryptographic sense can, for example, be

- factoring the product of two large primes. (numeric crypto, complexity-theoretic analysis)
- guessing a key chosen from a key-space which is as large as the message-space. (information-theoretic analysis)
- solving a problem where the AI-community agrees that it can easily be solved by intelligent humans, but that it cannot be solved within any known formal model. (HIPs)

Steganographic Security



Steganographic Security



Lexical Steganography

$C = \{$ Midshire is a nice little city,
Midshire is a fine little town,
Midshire is a great little town,
Midshire is a decent little town,
Midshire is a wonderful little town $\}$

$M = \{ 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 \}$

Lexical Steganography

Midshire is a {
wonderful
decent
fine
great
nice
} little {
city
town
} .

Lexical Steganography

Midshire is a $\left\{ \begin{array}{l} 00 \text{ wonderful} \\ 01 \text{ decent} \\ \mathbf{10} \text{ fine} \\ 11 \text{ great} \\ ?? \text{ nice} \end{array} \right\}$ little $\left\{ \begin{array}{l} 0 \text{ city} \\ \mathbf{1} \text{ town} \end{array} \right\}$.

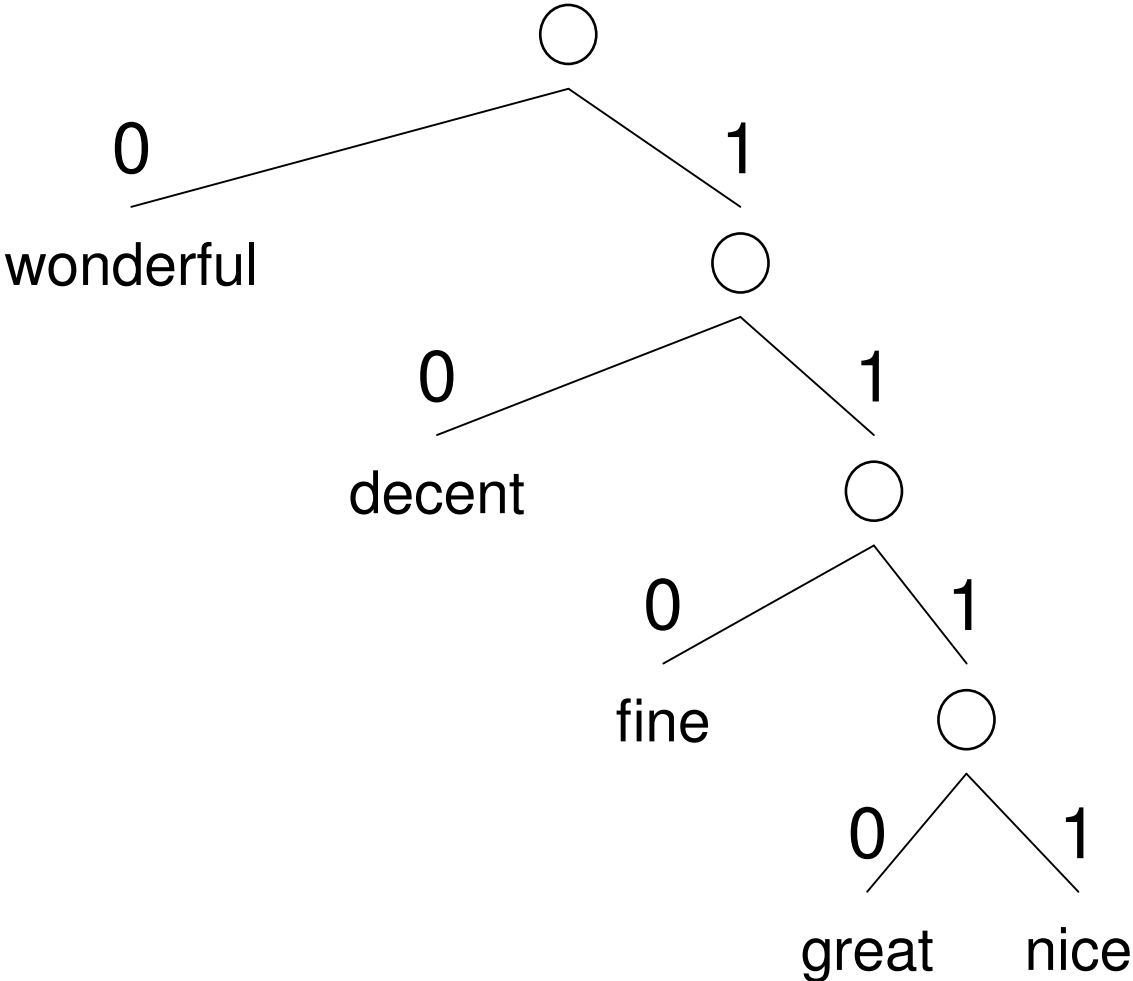
$$10|1 = 101_2 = \mathbf{5}_{10}$$

Lexical Steganography

Midshire is a $\left\{ \begin{array}{l} 0 \text{ wonderful} \\ 1 \text{ decent} \\ 2 \text{ **fine**} \\ 3 \text{ great} \\ 4 \text{ nice} \end{array} \right\}$ little $\left\{ \begin{array}{l} 0 \text{ city} \\ 1 \text{ **town**} \end{array} \right\}$.

$$\left[\begin{array}{l} 2, \\ 5, \end{array} \begin{array}{l} 1 \\ 2 \end{array} \right] = 2 * 2 + 1 = \mathbf{5}.$$

Lexical Steganography



Lexical Steganography

Midshire is a	0	wonderful	.5	} little	0	city	.5
	10	decent	.25		1	town	.5
	110	fine	.125				
	1110	great	.0625				
	1111	nice	.0625				

$$10|1 = 101_2 = \mathbf{5}_{10}$$

Lexical Steganography

All approaches we have seen so far have one basic idea in common: transforming a sequence of symbols

$$s_1, s_2, s_3, \dots, s_n$$

into a sequence

$$T(s_1) \mid T(s_2) \mid T(s_3) \mid \dots \mid T(s_n),$$

which has a “dual” interpretation, one with regard to the cover-channel, one with regard to a secret message.

Context-Free Mimicry

A more sophisticated linguistic model can be achieved, by assuming the symbols as grammatical productions

$$S \Rightarrow \alpha_1, \quad \alpha_1 \Rightarrow \alpha_2, \quad \alpha_2 \Rightarrow \alpha_3, \quad \dots, \quad \alpha_{m-1} \Rightarrow e.$$

into a sequence

$$T(S \Rightarrow \alpha_1) \mid T(\alpha_1 \Rightarrow \alpha_2) \mid T(\alpha_2 \Rightarrow \alpha_3) \mid \dots \mid T(\alpha_{m-1} \Rightarrow e)$$

which has a “dual” interpretation, one with regard to the cover-channel, one with regard to a secret message.

Chapman's system

The Doe and the Lion A DOE hard fixed by robbers taught refuge in a slave tinkling to a Lion. The Goods under- took themselves to aversion and disliked before a toothless wrestler on their words. The Sheep, much past his will, married her backward and forward for a long time, and at last said, If you had defended a dog in this wood, you would have had your straits from his sharp teeth. One day he ruined to see a Fellow, whose had smeared for its pro- vision, resigning along a fool and warning advisedly. said the Horse, if you really word me to be in good occasion, you could groom me less, and proceed me more. who have opened in that which I blamed a happy wine the horse of my possession.

[...]

Wayner's system

It's time for another game between the Whappers and the Blogs in scenic downtown Blovonia . I've just got to say that the Blog fans have come to support their team and rant and rave . Play Ball ! Time for another inning . The Whappers will be leading off . Baseball and Apple Pie . The pitcher spits. Herbert Herbertson swings the bat to get ready and enters the batter's box . Here's the fastball . He tries to bunt, and Robby Rawhide grabs it and tosses it to first . Hey, one down, two to go. Here we go. Prince Albert von Carmicheal swings the baseball bat to stretch and enters the batter's box . Okay. Here's the pitch It's a spitter . High and outside . Ball . No contact in Mudsville ! Nothing on that one . Nice hit into short left field for a dangerous double and the throw is into the umpire's head ! [...]

Winstein's system

“Risky E-Vote System to Expand” Wired News (01/26/04); Zetter, Kim [...]

She promises that the workplace computers people use to vote on SERVE will be

fortified⁽¹⁾ with firewalls and other intrusion countermeasures, and adds that election officials will recommend that home users install antivirus software on their PCs and run virus checks prior to election day.

Rubin counters that antivirus software can only identify known viruses, and thus is ineffective against new e-voting malware; moreover⁽¹⁾, attacks could go undetected because SERVE lacks elector⁽⁰⁾ verifiability.

Rubin and the three⁽¹⁾ other researchers who furnished the report were part of a 10-member expert panel enlisted by the Federal Voting Assistance Program (FVAP) to assess SERVE. Paquette reports that of the six remaining FVAP panel members, five recommended that the SERVE trial proceed, and one made no comment. [...]

{bastioned⁽⁰⁾, fortified⁽¹⁾}, {furthermore⁽⁰⁾, moreover⁽¹⁾}, {elector⁽⁰⁾, voter⁽¹⁾}, {iii⁽⁰⁾, three⁽¹⁾}

For a number of reasons, I believe that the basic approach that is most promising for building a secure and robust natural language steganography system in the near future is the **lexical replacement system, similar in principle to Winstein's.**

The state of the art in computational linguistics and artificial intelligence is a significant limiting factor!

- Do ontological semantics scale?
- Even if they did, we do not have a reliable common-sense ontology, yet.
- Context-free grammars alone do not adequately characterize natural languages. ($a^n b^n c^n$ respectively)
- Style-templates were never meant to fool sophisticated linguistic models or humans.

- Lexical models do scale!
- And we even have large-scale resources available, that cover all of everyday written language. (WordNet, for instance)
- Lexical models do not dig very deep into the semantic realm, but usually this will not be a problem, if
- we use an embedding-approach, instead of a generation-approach. This rather conservative approach follows the policy: “Use human language-competence as much as possible, and rely on formal models only when necessary!”

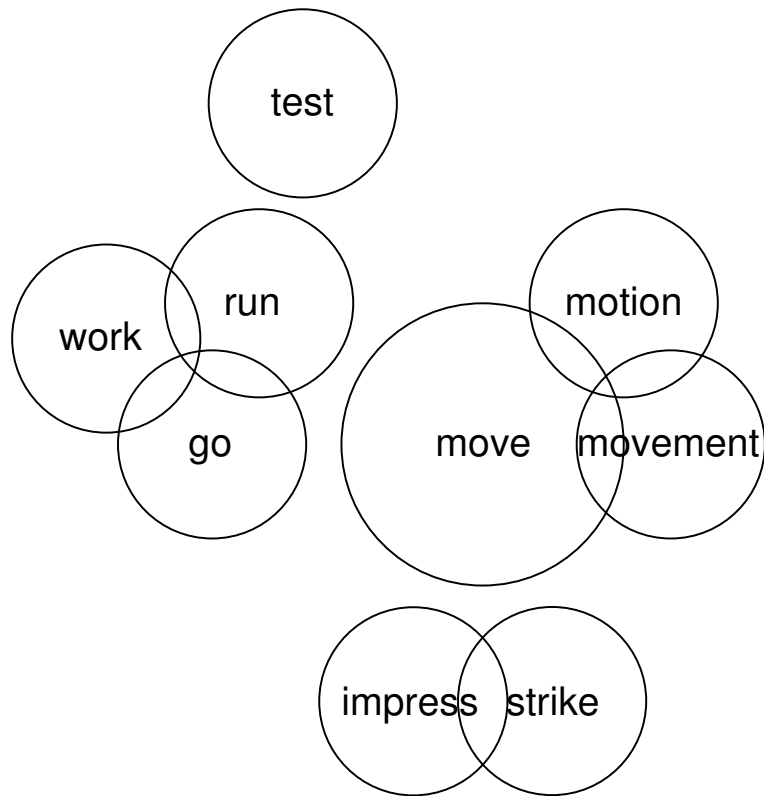
current systems

- do not mimic cover-statistics adequately: They do not mimic word-choice probabilities. A system similar in principle to Winstein's, however following Wayner's coding strategy, should be used instead.
- do not encrypt messages adequately: Everyone can extract the messages from the steganograms if he has the correct dictionary, respectively grammar. (Shouldn't linguistic knowledge be assumed public wisdom? Language is, by definition, something public!) Messages should be encrypted with respect to key-distribution systems instead!

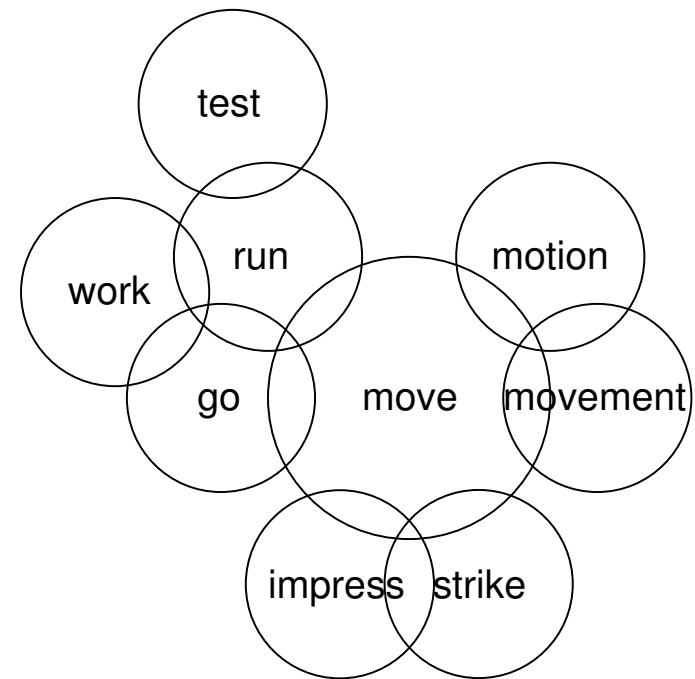
current systems

- lack robustness. Some kind of error-correction should be applied.
- employ linguistically inadequate models: They use disjunct interchangeability sets. Statistical word-sense disambiguation systems should be used instead.

Lexical Ambiguity and Coding

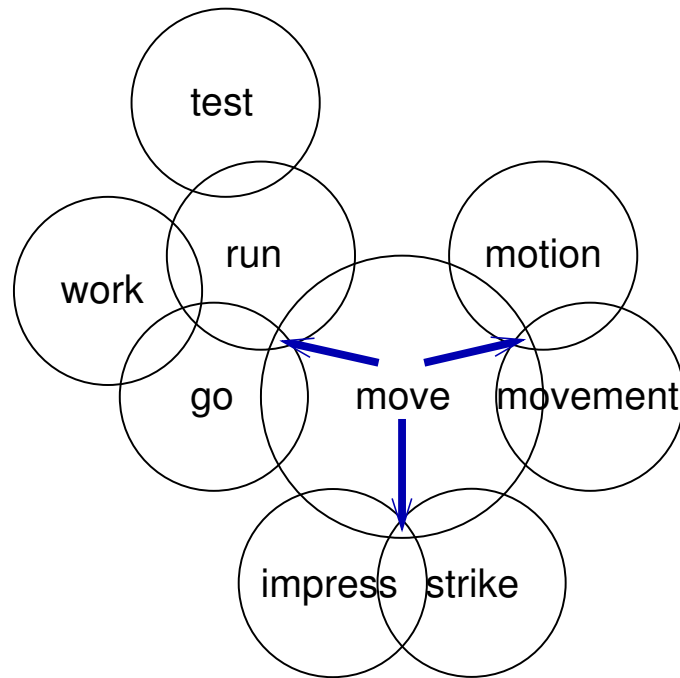


(a) disjunct synsets

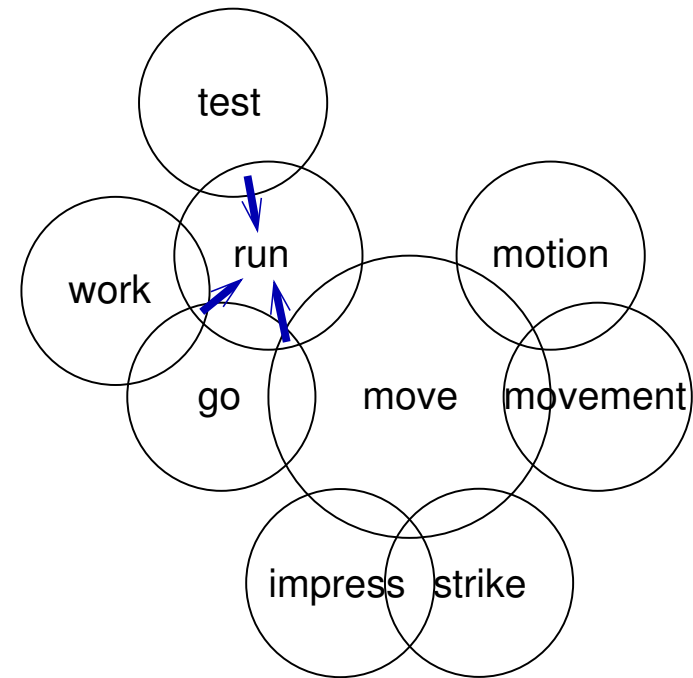


(b) natural synsets

Lexical Ambiguity and Coding

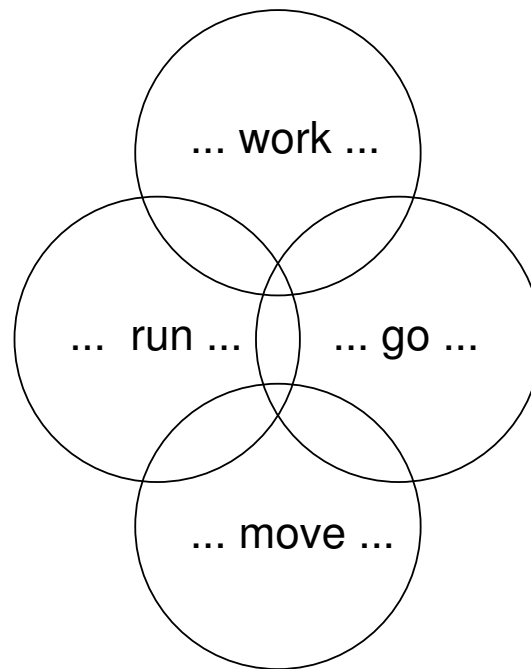


(c) "Forward ambiguity"

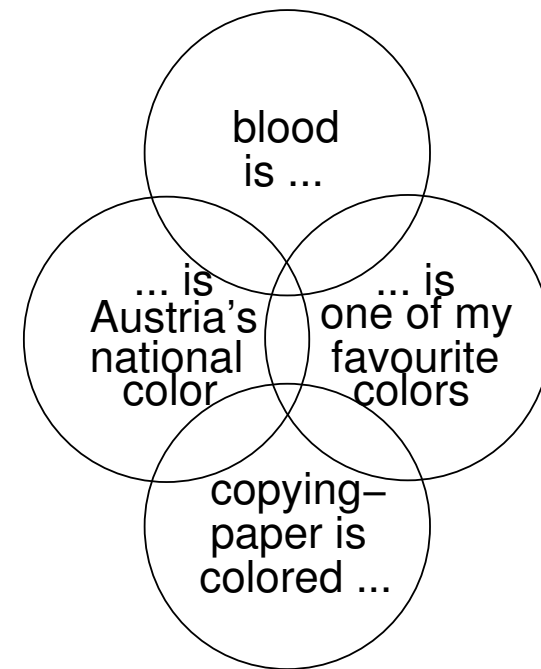


(d) "Backward ambiguity"

Lexical Ambiguity and Coding



(e) lexical semantics



(f) "contextual" semantics

Lexical Ambiguity and Coding

Uncle Joe turned out to be a brilliant player of the electric guitar.

$\mathcal{C}(\textit{brilliant}) = \langle \text{Joe, turned, } \textit{brilliant}, \text{player, electric} \rangle,$

$\mathcal{C}(w) = \langle w_{-3}, w_{-2}, w_{-1}, w_0, w_1, w_2, w_3 \rangle,$

$$P(\mathcal{C}(x)|s) = \prod_{j=-n}^n P(w_j|s),$$

$$P(s|\mathcal{C}(w)) = \frac{P(s)P(\mathcal{C}(w)|s)}{P(\mathcal{C}(w))}.$$

Lexical Ambiguity and Coding

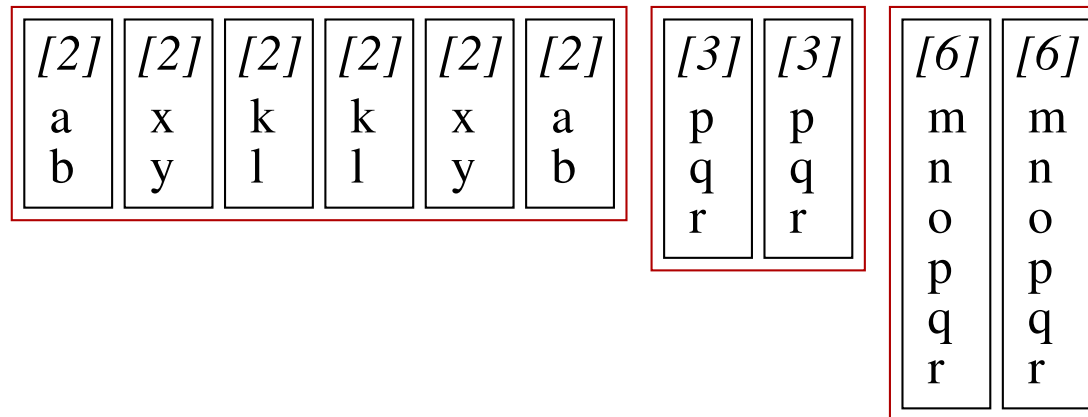
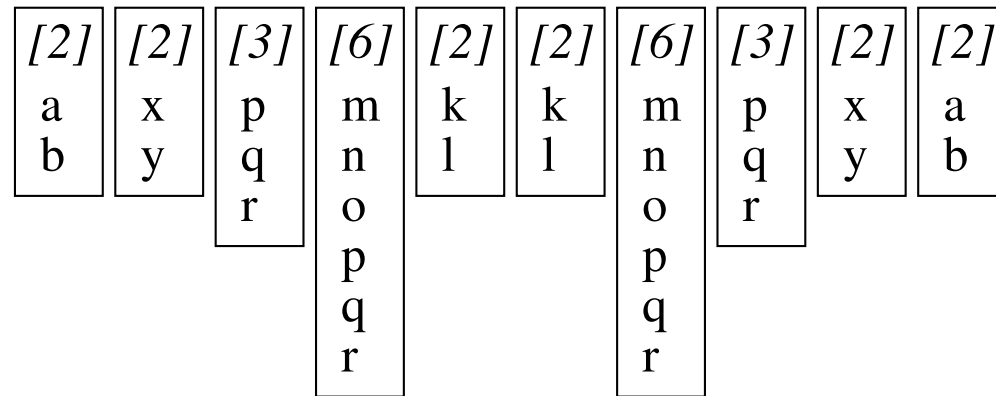
$$\text{rep}(o) = \text{dis}(\mathcal{L}(o), \mathcal{C}(o)).$$

$$r \in \text{rep}(o) \Rightarrow r \in \begin{cases} \text{rep}_A(o), & \text{if } \text{rep}(o) = \text{rep}(r) \\ \text{rep}_B(o), & \text{otherwise.} \end{cases}$$

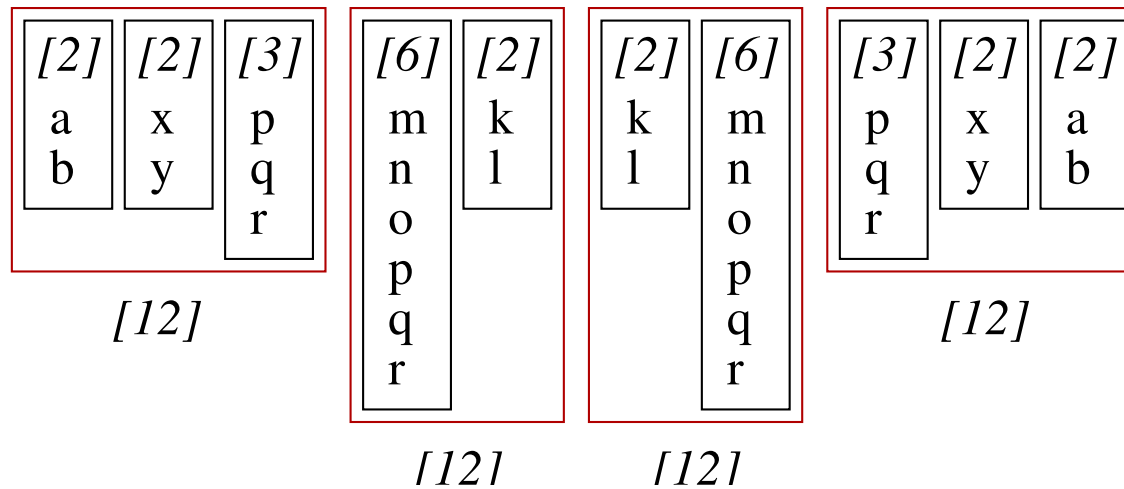
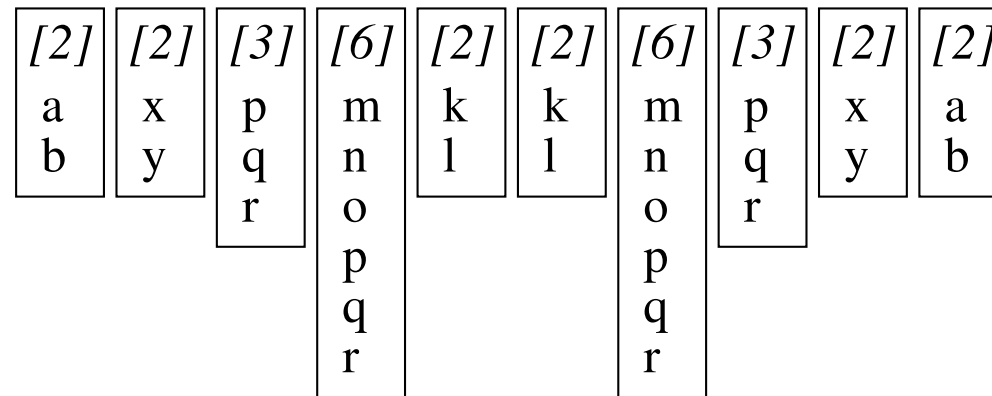
Lexical Ambiguity and Coding

- type-A-words o where $\text{rep}_B(o) = \emptyset$. Here we can be sure that a replacement of word o will *always* be reversible automatically.
- type-B-words o where $\text{rep}_A(o) = \emptyset$. Here we can be sure that a replacement of word o will *never* be reversible automatically.
- type-C-words o where $\text{rep}_A(o) \neq \emptyset \wedge \text{rep}_B(o) \neq \emptyset$. Here the question whether a replacement will be reversible depends on the actual replacement which is made.

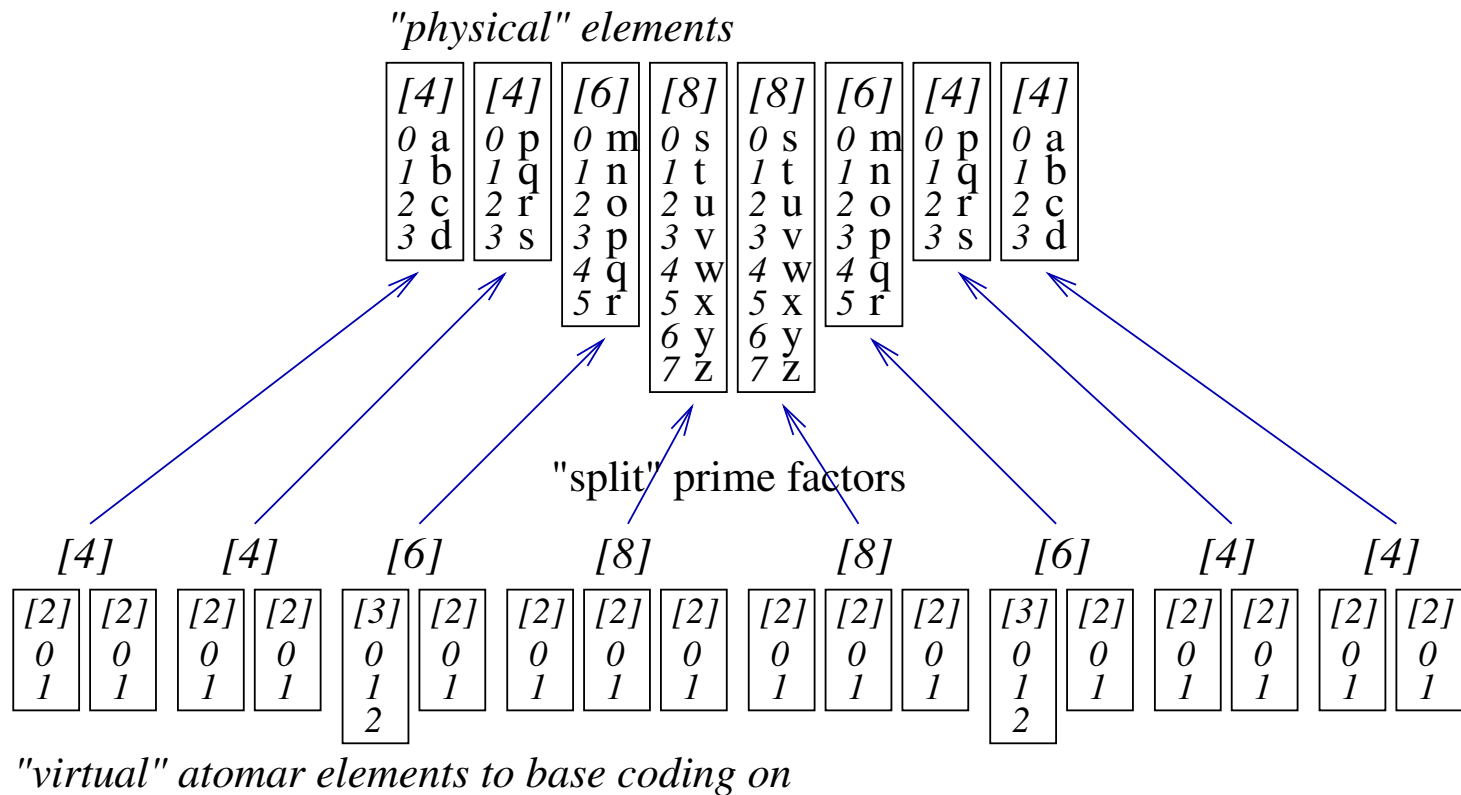
Secure and Robust Coding



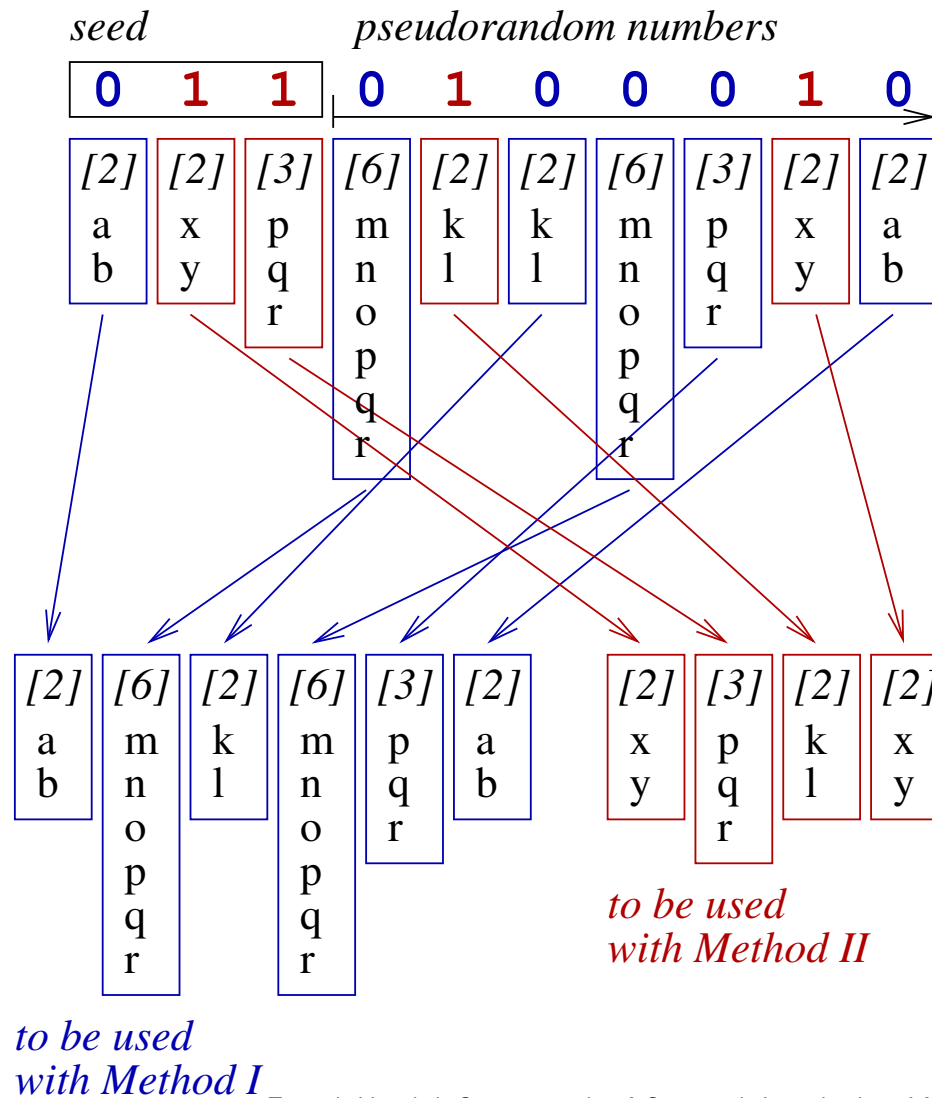
Secure and Robust Coding



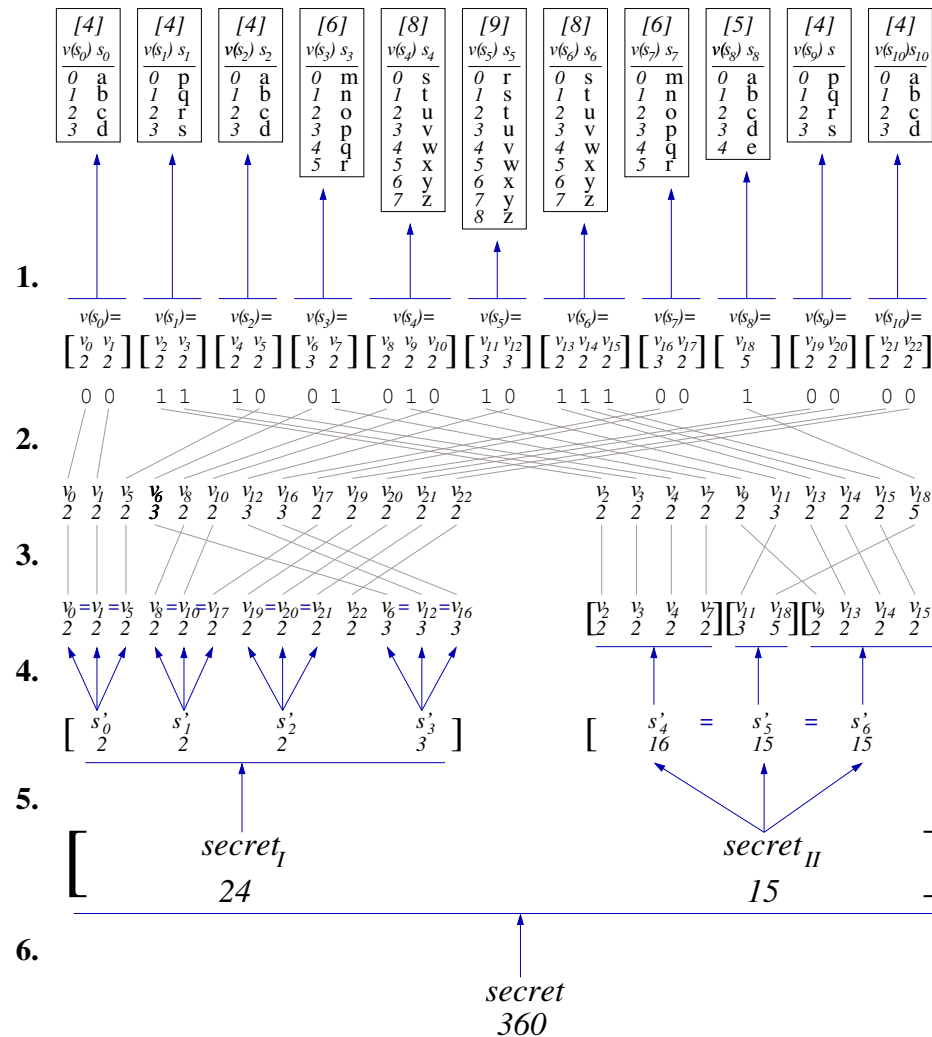
Secure and Robust Coding



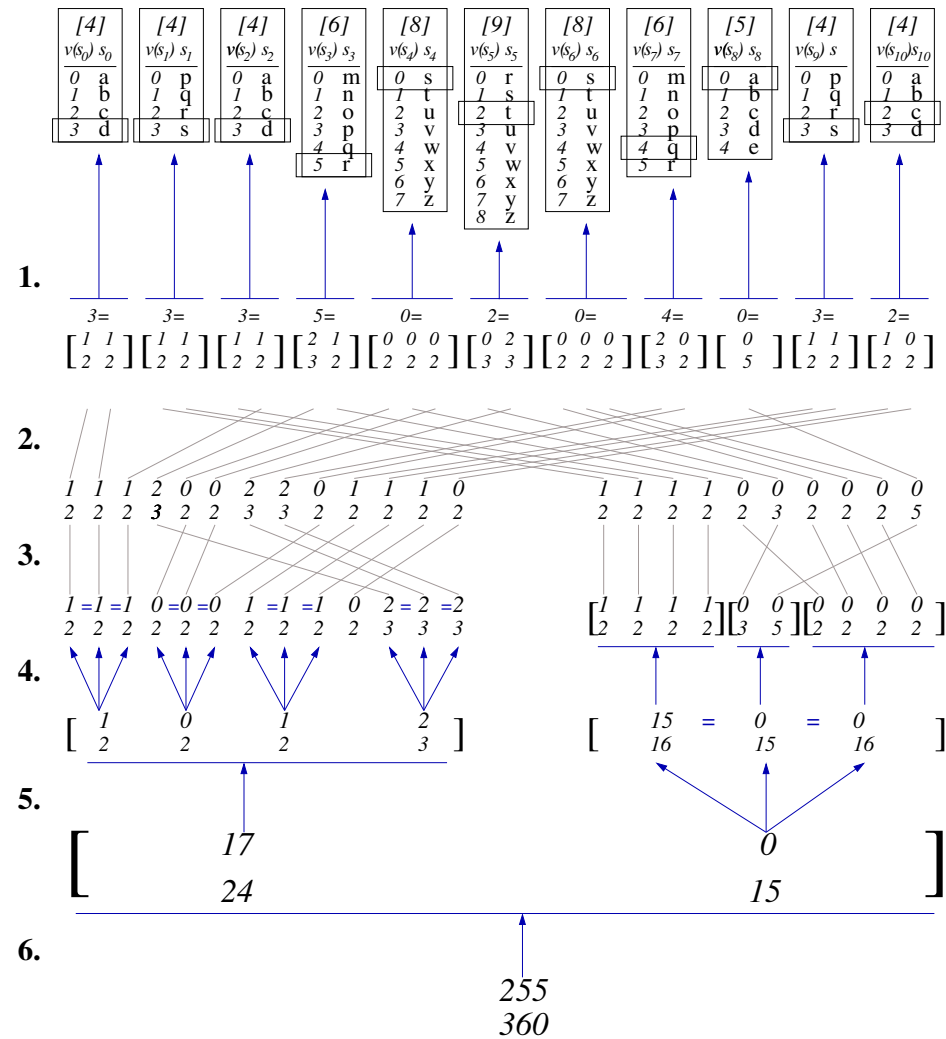
Secure and Robust Coding



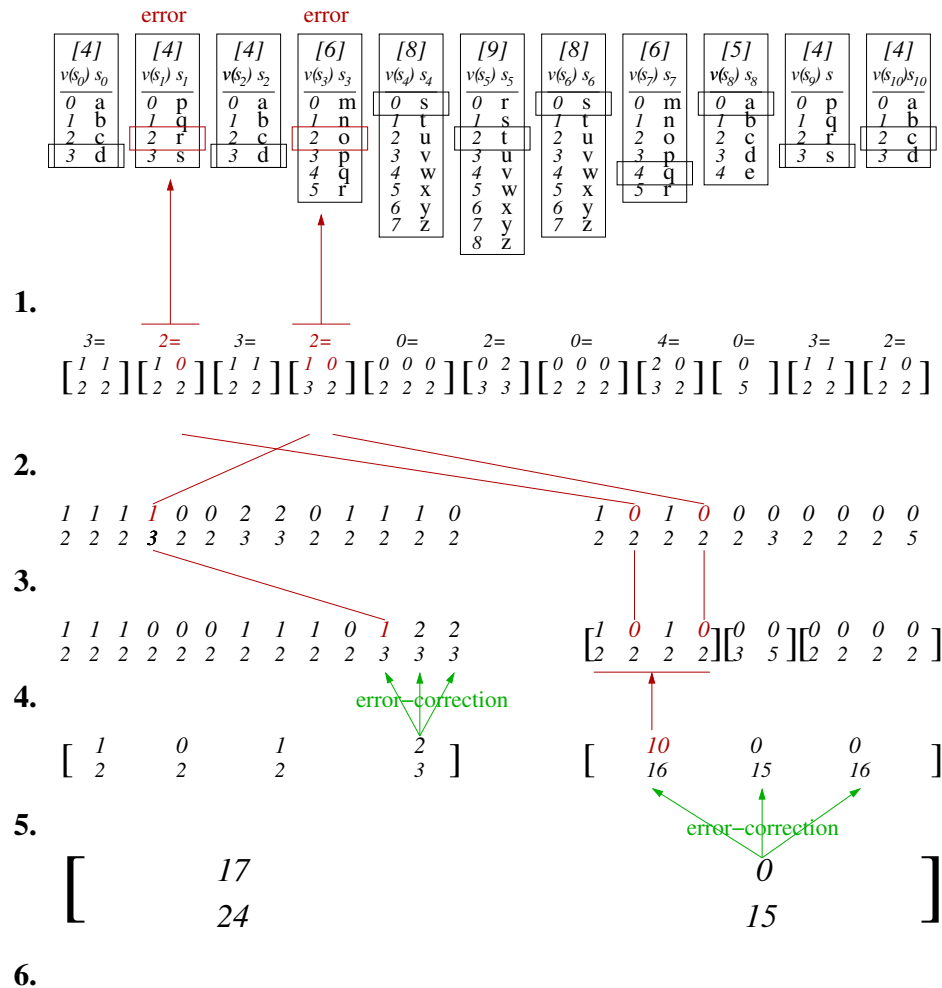
Secure and Robust Coding



Secure and Robust Coding



Secure and Robust Coding



255
360



Towards Linguistic Steganography:
A Systematic Investigation of Approaches, Systems,
and Issues

A project conducted Oct-03 – Apr-04 by

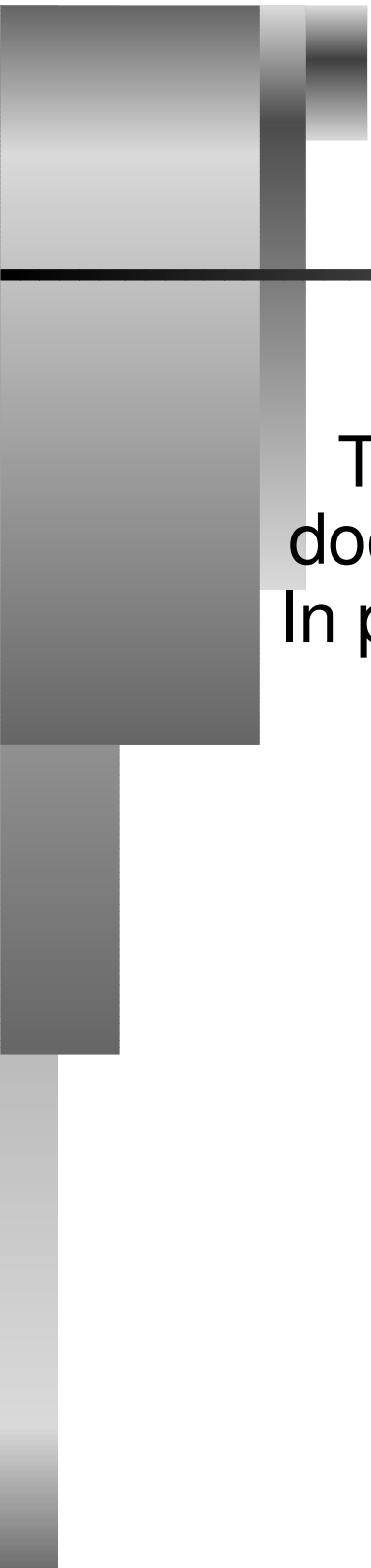
Richard Bergmair

at

University of Derby in Austria

under supervision by

Stefan Katzenbeisser.



This slide-set is not to be seen as a self-contained document. Please conduct the project-report instead. In particular, note that sources were not properly cited in this slide-set. See the citations given in the project-report for reference on sources.