Методы сокрытия информации в текстовых документах

Метод	Описание
Изменение количества пробелов	Данный метод имеет несколько разновидностей. 1. Внедряемая информация кодируется одним или двумя пробелами в конце предложения. 2. Для сокрытия информации один или два пробела добавляются в конце каждой строки. 3. При внедрении между каждым словом контейнера ставится один или два пробела. 4. Кодирование осуществляют путем формирования четного или нечетного числа символов в строке.
Синтаксический метод	Метод позволяет скрывать сообщения путем изменения пунктуации.
Орфографический метод	Скрытое сообщение внедряется с помощью намеренно сделанных орфографических ошибок.
Семантический метод	Метод базируется на кодировании информации с помощью словаря синонимов.

Изменение формата текста	1. Изменение кернинга (расстояния между соседними буквами). 2. Изменение ширины отступов («красной» строки). 3. Изменение начертания букв (например, использование для кодирования латиницы и кириллицы или изменение гарнитуры). 4. Изменение интерлиньяжа (расстояния между строками). 5. Изменение цвета букв. 6. Изменение размеров символов (например, точек).
Внедрение информации в поле комментариев	Информация скрывается с помощью непечатаемых символов в поле комментариев архива.

Лекция 4. Стегосистемы для других покрывающих сообщений.

4.1. Лингвистические СГС (Л-СГС)
4.2. Графические СГС (Г-СГС)
4.3. Интернет СГС (И-СГС)

4.1. Лингвистическая стегосистема.

Определение: Скрытое вложение любых оцифрованных данных в файлы текстовых документов, использующих естественные языки.

Основные требования: Л-СГС не должна вызывать подозрений, т.е. вся структура языка (грамматика, синтаксис, семантика) должна сохраняться.

Два основных типа Л-СГС:

1. С заданным ПС (текстом).
2. С выбираемым ПС.

Основной принцип построения Л-СГС 1го типа.

Находить участки равномерно распределенные в некоторой области и заменять их другими по правилам вложения секретной информации.

Основной метод построения Л-СГС – использование абсолютных или относительных *синонимов*.

Определения:
1. Абсолютный синоним – это слово или фраза, которые могут быть заменены другим словом или фразой в любом контексте без изменения его смысла.

Примеры наборов абсолютных односложных синонимов:

взгляд – взор, годный пригодный, гостиница – отель, громадный – огромный, грусть – печаль, отличник – пятерочник, доля – часть, заглавие – заголовок, заграничный – зарубежный, зыбь – рябь, иссяк – истощился, показалось – почудилось, лгун – лжец, многократно – неоднократно, незаконный – противозаконный, и др.

Примеры синонимов на английском языке:

sofa – settee, big – large, another – different, mind – opinion, and so on.

Примеры синонимов – фраз (в том числе и сокращения):

Соединенные Штаты – США, бывший президент – экс-президент, центральная избирательная комиссия – ЦИК, и т.д.

2. *Относительные синонимы* это слова или фразы, которые могут, заменить друг друга (или нет) в зависимости от контекста (окружения этих слов или фраз).

Примеры относительных синонимов:

дать ход (документу) – направить,

дать ход (от преследователей) – уехать.

Примеры относительных синонимов на английском языке:

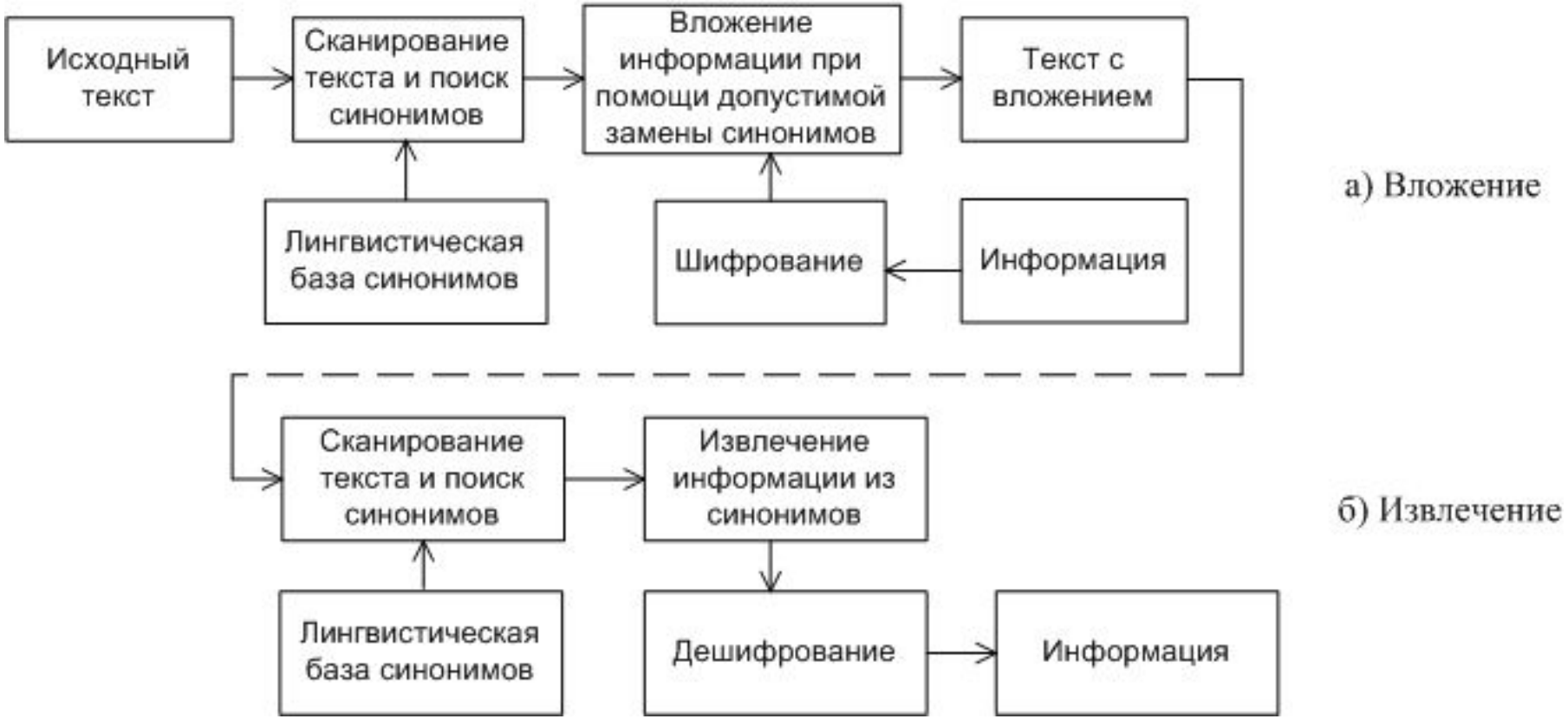
real number – continuous number, real life ≠ continuous life.

Абсолютные и относительные синонимы для каждого языка собраны в специальные словари, например:

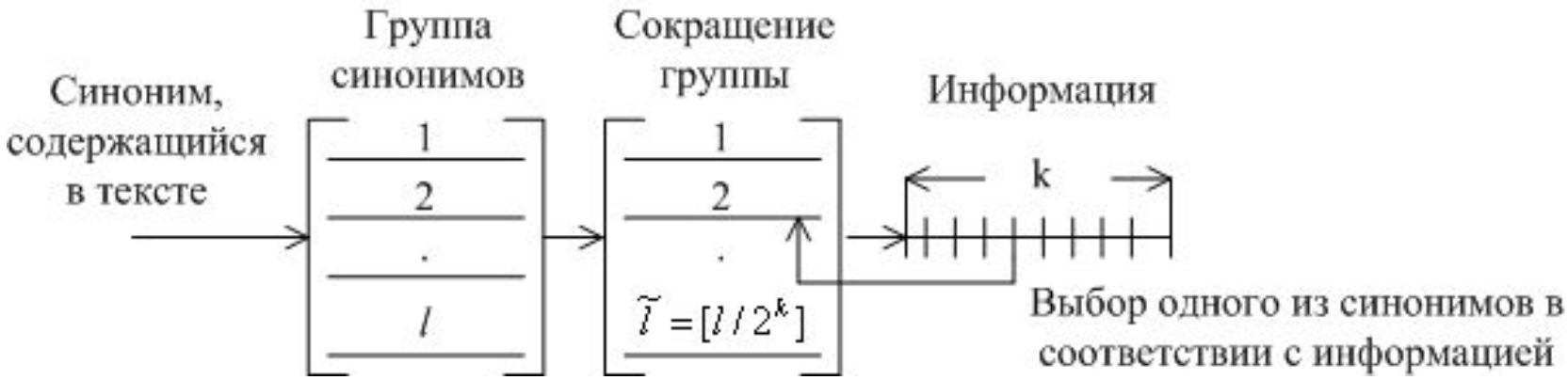
1. Словарь синонимов русского языка.
2. Oxford Collocation Dictionary for Students of English. Oxford University Press. 2003.
3. Fellbaum Ch. WordNet: An electronic lexical database. MIT Press, 1998.

Замечание. Важное понятие “collocation” (взаиморасположение) – это допустимое сочетание соседних слов.

Алгоритм вложения и извлечения секретной информации для Л-СГС.



Метод вложения информации в группу синонимов.



Примеры:

---------- → 00 ---------- → 0

---------- → 01 ---------- → 1

---------- → 10

---------- → 11

(Возможны и более эффективные методы кодирования).

Пример построения Л-СГС.

Исходный текст.

Пять подземных толчков *зарегистрировано* за сутки на юге *республики Алтай. Сила* землетрясений *составила* от 2.2 до 3.1 балла по шкале Рихтера, *сообщили* на Алтайской сейсмической станции сегодня *после полуночи*.

(Абсолютные синонимы подчеркнуты сплошной линией, а относительные – выделены курсивом).

Группы абсолютных синонимов:

землетрясение (0) – подземные толчки (1),
за сутки (0) – за 24 часа (1),
сейсмическая станция (0) – сейсмостанция (1).

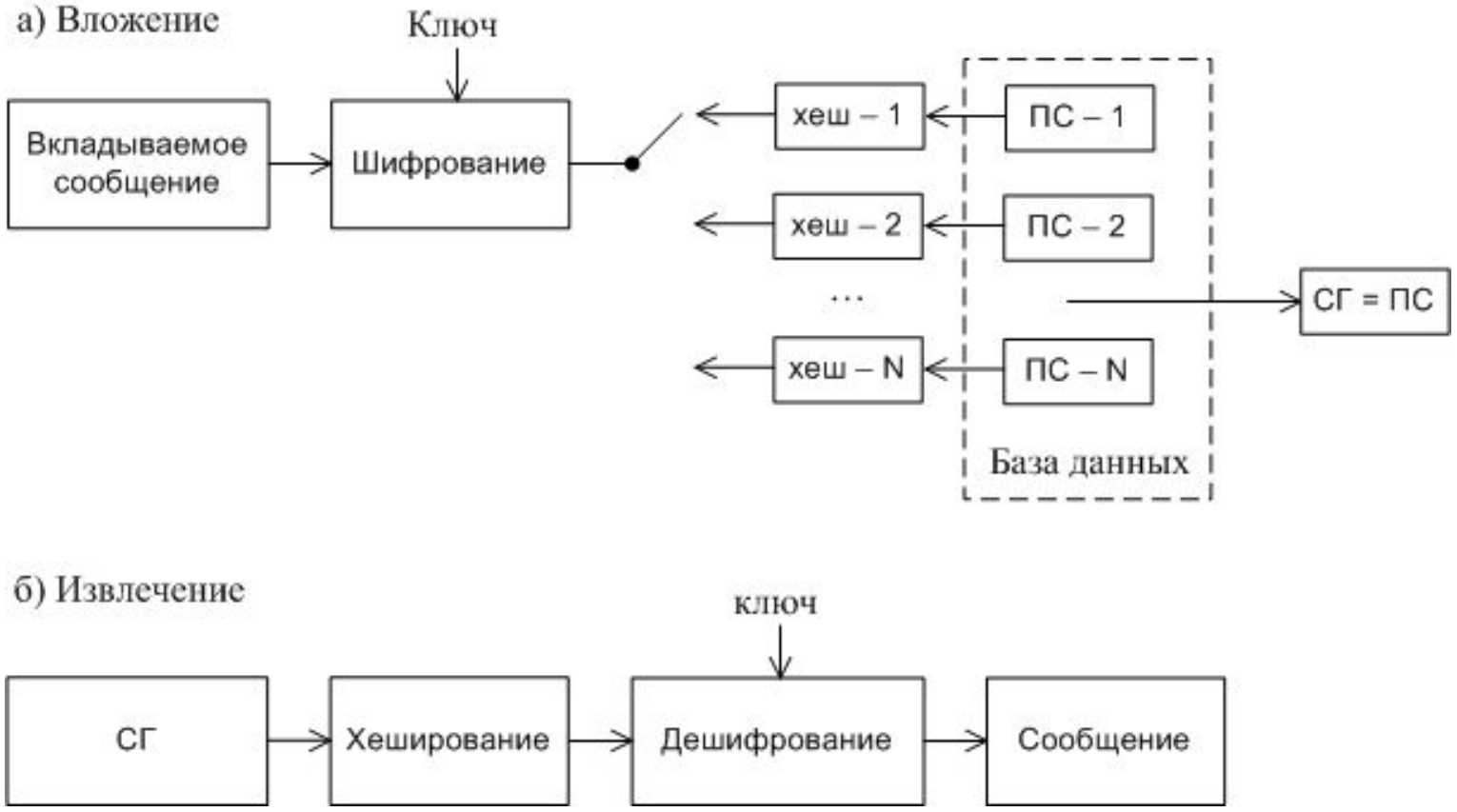
Группы относительных синонимов:

зарегистрированный (00),
зафиксированный (01),
отмеченный (10),
замеченных (11),
Республика Алтай (0) – Алтай (1),
составила (0) – равнялась (1),
проинформировать (0) – сообщить (1),
после полудня (0) – во второй половине дня (1),
сила (00), амплитуда (01), мощь (10), мощность (11).

После проверки групп относительных синонимов на совместимость с их “окружением”, производится выбор замен по заданной информации.
Общее количество вложенных бит равно 12.
(В данном фрагменте можно передать 2 латинские буквы в коде ASCII. Это примерно 0.73% от объема ПС.)
Другой метод Л-СГС: Изменение порядка слов в предложении.
Пример: В Иране во вторник произошло новое землетрясение.
L T V S
L – обстоятельство места, *T* – обстоятельство времени, *V* – сказуемое, *S* – подлежащее.
Всего возможно 4! = 24 перестановки, но абсолютно допустимо 3:
TLVS, SVTL, TVSL – LTVS (исходное).
Еще возможно 10 вариантов, но с другими оттенками, например *VTLS* (“Произошло во вторник в Иране новое землетрясение”…).
Количество вкладываемых бит: 2 – 3.

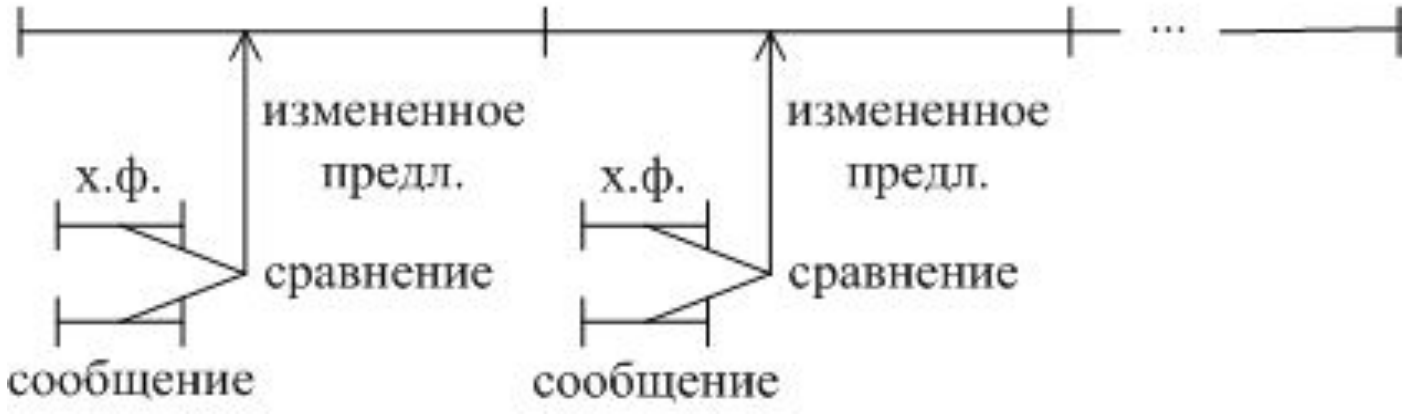
Сложность данного метода – в невозможности его автоматизации.
(Автоматизированный разбор предложения – не решенная пока проблема *структурной лингвистики*.)
Еще один метод Л-СГС – изменение шрифта (Шекспир) и использование совпадающих букв в русском и английском.

2. Л-СГС с выбираемыми (конструируемыми) ПС.
Это частный случай общего метода построения идеальных СГС:



Замечание. В этой системе хеш-функция и база данных общеизвестны, а секретен только ключ шифрования / дешифрования, который совпадает со стегоключом.

Обобщение на случай Л-СГС:



Свойства всех Л-СГС:

1. Идеальная секретность.
2. В качестве ПС может выступать любой смысловой текст.
3. Низкая скорость вложения.
4. Отсутствует устойчивость к “слепой” атаке удаления вложенной информации.
5. Иногда требует участия человека – оператора.

4.2. Графические СГС.

ПС – графический (растровый) документ (текст, картинка, схема, формула, и т. п.)

Простейшие методы погружения в графические текстовые документы:

- изменение расстояний между словами и предложениями,
- изменение пробелов между строками,
- сдвиги слов вверх и вниз,
- небольшие вращения строк.

(См. демонстрацию на следующем слайде.)

Все эти СГС легко обнаруживаются при использовании статистического стегоанализа.

Примеры вложения информации в текстовые файлы:

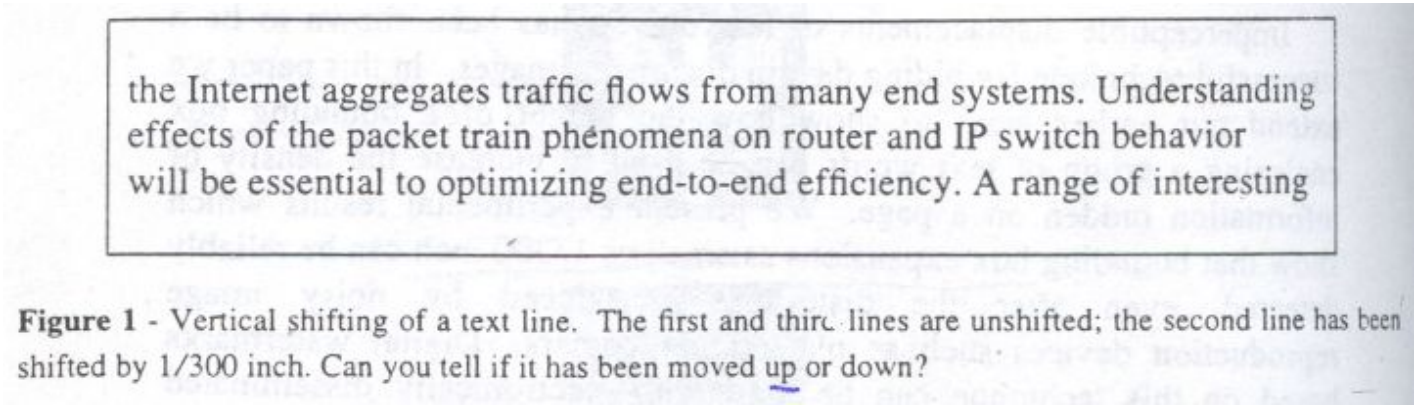


Figure 1 - Vertical shifting of a text line. The first and third lines are unshifted; the second line has been shifted by 1/300 inch. Can you tell if it has been moved up or down?

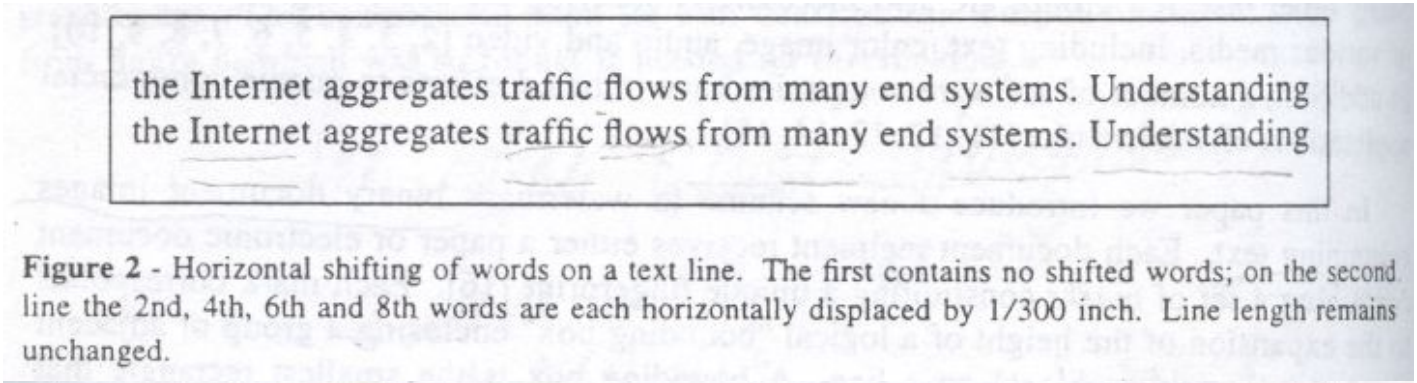


Figure 2 - Horizontal shifting of words on a text line. The first contains no shifted words; on the second line the 2nd, 4th, 6th and 8th words are each horizontally displaced by 1/300 inch. Line length remains unchanged.

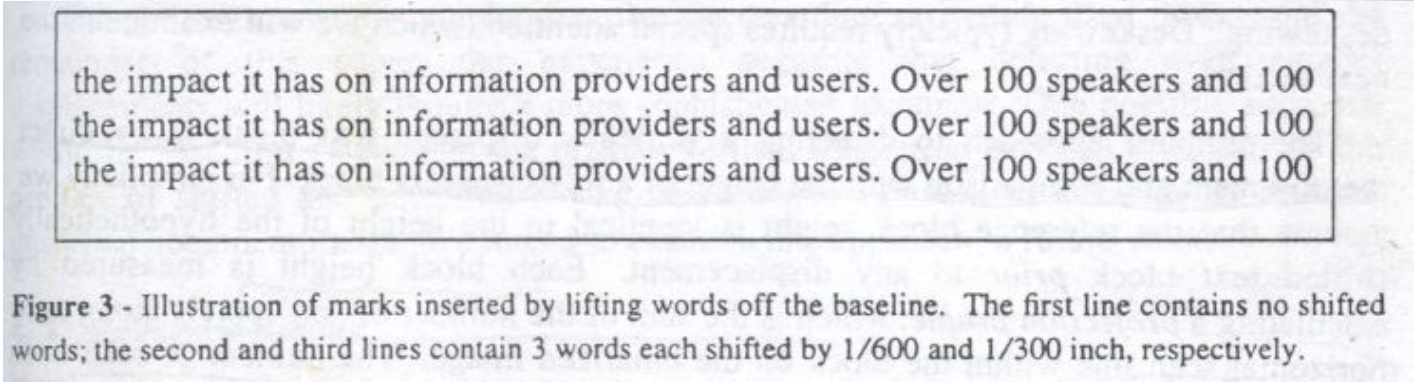


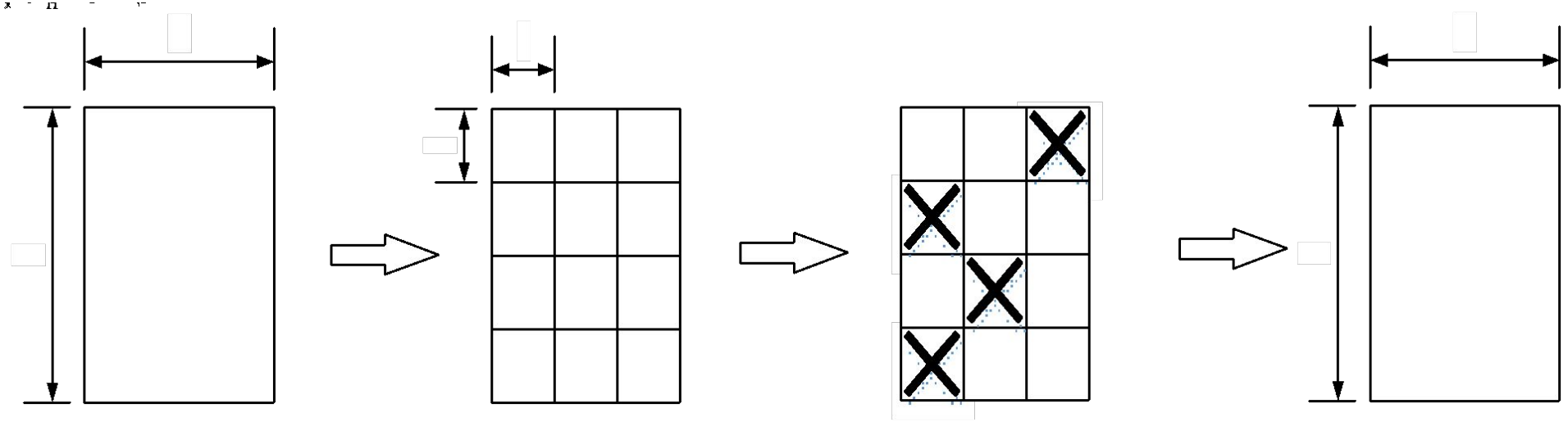
Figure 3 - Illustration of marks inserted by lifting words off the baseline. The first line contains no shifted words; the second and third lines contain 3 words each shifted by 1/600 and 1/300 inch, respectively.

Более изощренный метод “Имитация шумов сканирования”.

Основная идея: Отсканировать напечатанный документ и внести в него скрытную информацию, имитируя шумы сканера.

Метод погружения скрытой информации:

1. Отсканированный черно-белый документ последовательно делится на области *nxn* пикселей *A*.



Введем обозначения: m – количество черных пикселей в *A, m_+* - количество черных пикселей в *A*, если оно четное, *m_-* - если нечетное, *$0 < k < ½$* выбранный порог, *b = {0,1}* значение бита скрытой информации, вкладываемой в *A, $A = A_0$,* если *$kn^2 < m < (1-k)n^2$, $A = A_1$,* если *$m = (1-k)n^2$*, *$A = A_2$,* если *$m = kn^2$*.

2. Если $A = A_0$, то вложение производится в соответствии с таблицей:

	$A = A_+$	$A = A_-$
b = 0	Ничего не изменять	Изменить цвет одного пикселя на противоположный
b = 1	Изменить цвет одного пикселя на противоположный	Ничего не изменять

Замечание. Изменяться могут любые пиксели, но только на границе черного и белого.

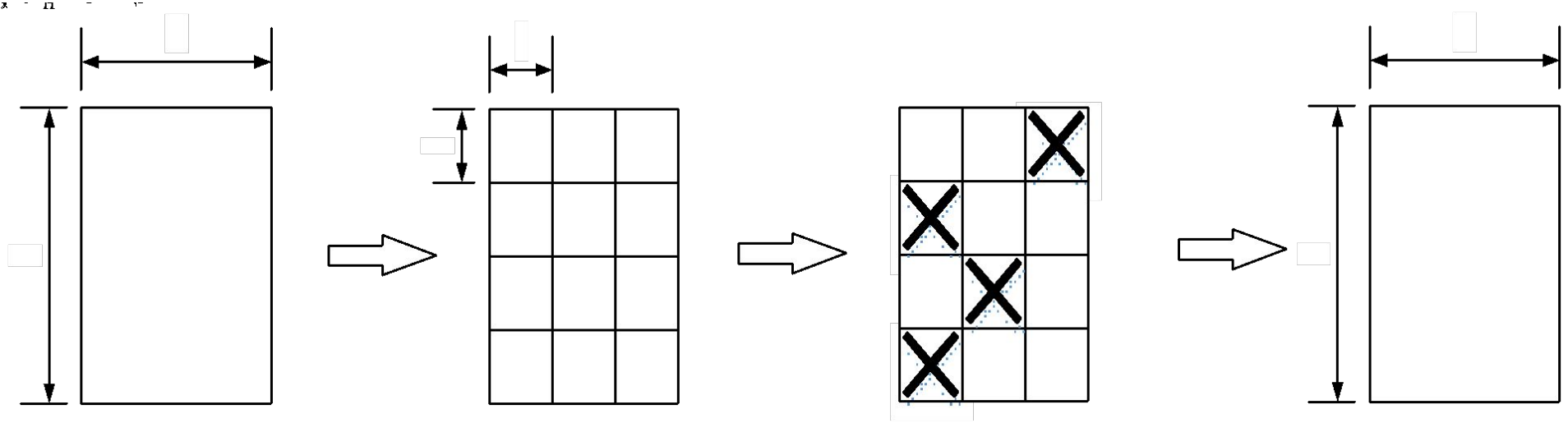
3. Если $A = A_1$, то вложение производится в соответствии с таблицей:

	$A = A_+$	$A = A_-$
b = 0	Ничего не изменять	Изменить один черный пиксель на белый
b = 1	Изменить один черный пиксель на белый	Ничего не изменять

4. Если $A = A_2$, то вложение производится в соответствии с таблицей:

	$A = A_+$	$A = A_-$
b = 0	Ничего не изменять	Изменить один белый пиксель на черный
b = 1	Изменить один белый пиксель на черный	Ничего не изменять

5. Если $A \neq A_0$, $A \neq A_1$, $A \neq A_2$ то ничего не вкладывать в эту область.



Метод извлечения скрытой информации:

1. Последовательно разделить изображение на *A*-области размером *nxn*.
2. Если $A=A_0$, или $A=A_1$, или $A=A_2$, то извлечь *b=0*, если $A=A_+$ и *b=1*, если $A=A_-$.
3. Если $A\neq A_0$, $A\neq A_1$, $A\neq A_2$, то не извлекать из этой области никакой информации.

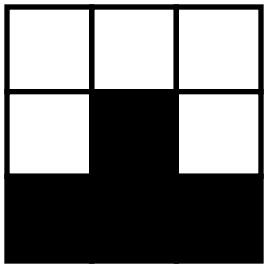
Основные свойства данного метода:

1. Извлечение информации производится без ошибок.
2. Чем больше *n* и чем больше *k*, тем секретнее вложение, но тем меньше скорость вложения, и наоборот.
3. Вложение устойчиво к визуальной атаке и к простейшим статистическим атакам.
4. Вложение легко удаляется при помощи рандомизации A_+, A_- без ухудшения качества документа.
5. Скорость вложения невелика.

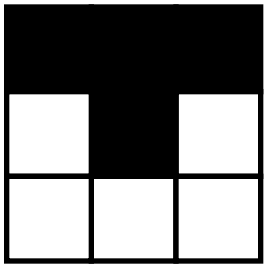
Атака, основанная на подсчете одиночных отклонений

Одиночные отклонений

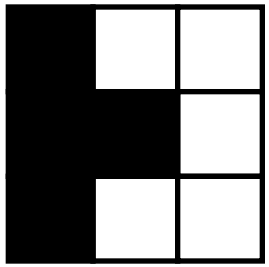
выброс



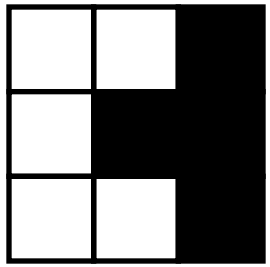
Up



Down

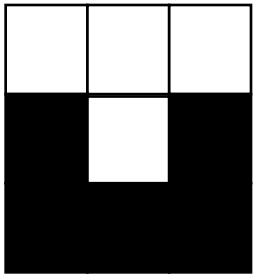


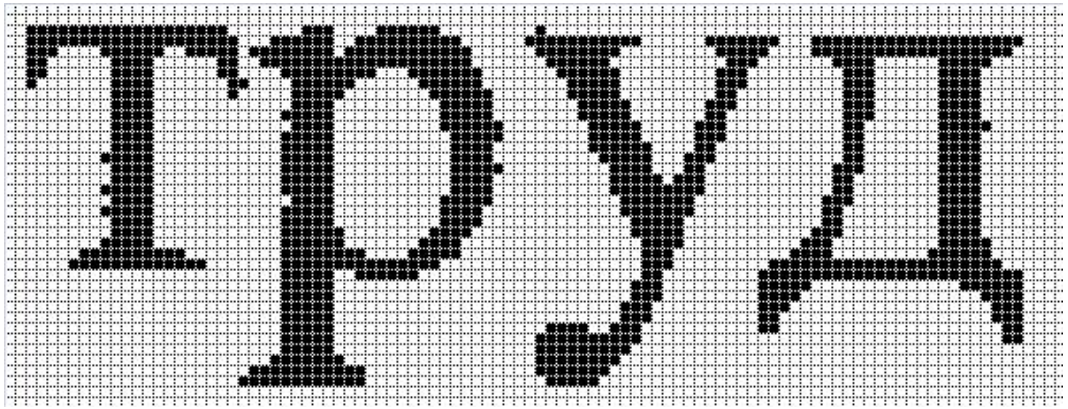
Right



Left

углубление





Атака, основанная на количестве одиночных отклонений

Гипотеза : тот же объем текста на странице формата А4 в среднем имеют меньшее число одиночных отклонений, чем после встраивания

Пороговое значение : количество одиночных отклонений для различных объемов текста

Анализ количества одиночных отклонений

Img. №	Up	Down	Left	Right	All before	Up	Down	Left	Right	All after
1	462	509	495	492	1958	602	542	649	720	2513
2	545	609	607	672	2433	654	654	757	936	3001
3	601	660	695	555	2511	743	713	866	787	3109
4	637	701	694	712	2744	788	773	851	951	3363
5	617	717	623	673	2630	799	791	806	887	3283
6	661	704	625	587	2577	818	770	787	825	3200
7	607	678	725	651	2661	750	735	862	903	3250
8	594	791	675	660	2720	743	866	819	897	3325
9	586	671	725	663	2645	728	728	897	881	3234
10	554	632	616	592	2394	691	708	761	815	2975
11	612	772	781	680	2845	782	830	937	915	3464
12	560	676	625	608	2469	696	726	788	823	3033
13	627	721	672	670	2690	746	776	782	885	3189
14	616	721	667	666	2670	780	767	845	872	3264
15	444	559	512	465	1980	565	622	621	666	2474
16	560	649	607	539	2355	693	695	778	764	2930
17	603	595	644	601	2443	734	655	788	808	2985
18	511	652	531	531	2225	591	705	682	742	2720
19	533	721	587	564	2405	665	782	748	790	2985
20	537	661	565	540	2303	679	716	702	792	2889

Атака, основанная на количестве одиночных отклонений

Ограничения для применения:

- Все текстовые документы печатаются на одном принтере;
- Все печатные документы сканируются на том же сканере;
- Необходима база данных тестовых изображений для сбора статистики;

Алгоритм обнаружения:

- Пороговые значения выбираются на основе собранной статистики в зависимости от распределения текста на странице;

  В качестве критерия для определения распределения текста на странице используется количество черных пикселей на странице.

- Поиск и подсчет единичных отклонений в отсканированный текстовом документе;
- Подсчет количества черных пикселей в отсканированном документе;
- Сравниваем подсчитанные единичные отклонения с выбранным пороговым значением;
- Принимается решение, является ли изображение ПО или СГ

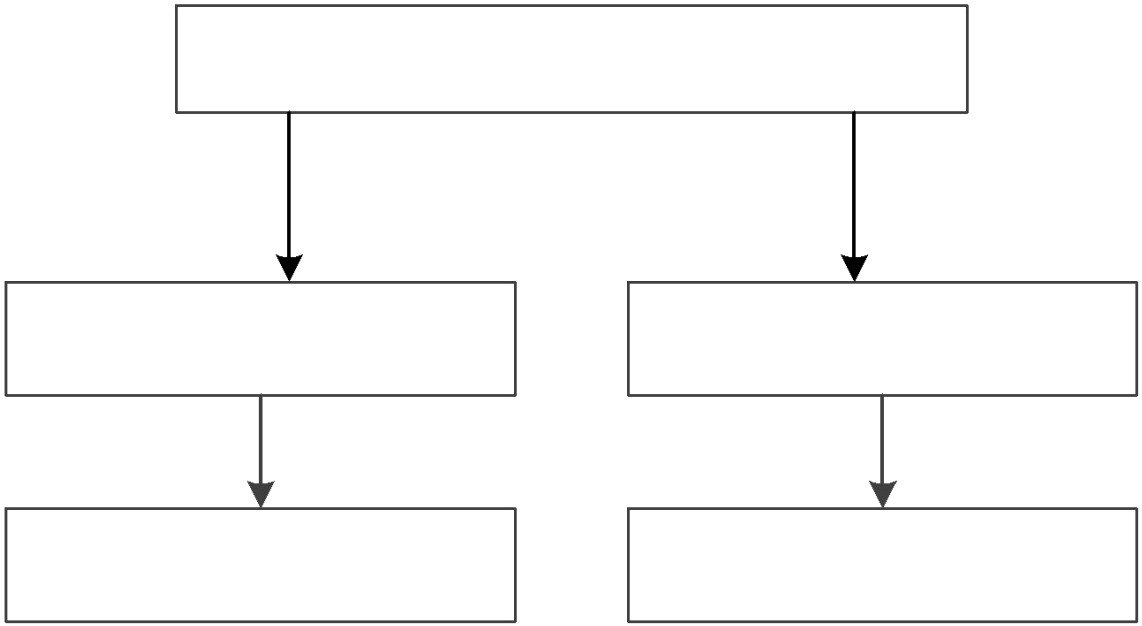
Оценка эффективности атаки, основанной на количестве одиночных отклонений

1. Следующие пороговые значения выбраны на основе анализа 20 тестовых изображений.
2. Скрытая информация внедряется с разной скоростью встраивания в 15 из 60 фотографий, представленных для стегоанализа

Количество черных пикс	Выбранный порог
600000 – 650000	1950
650000 – 700000	2150
700000 – 750000	2350
750000 – 800000	2550
800000 – 850000	2750
850000 – 900000	2950
900000 – 950000	3150

Оценка эффективности атаки, основанной на количестве одиночных отклонений

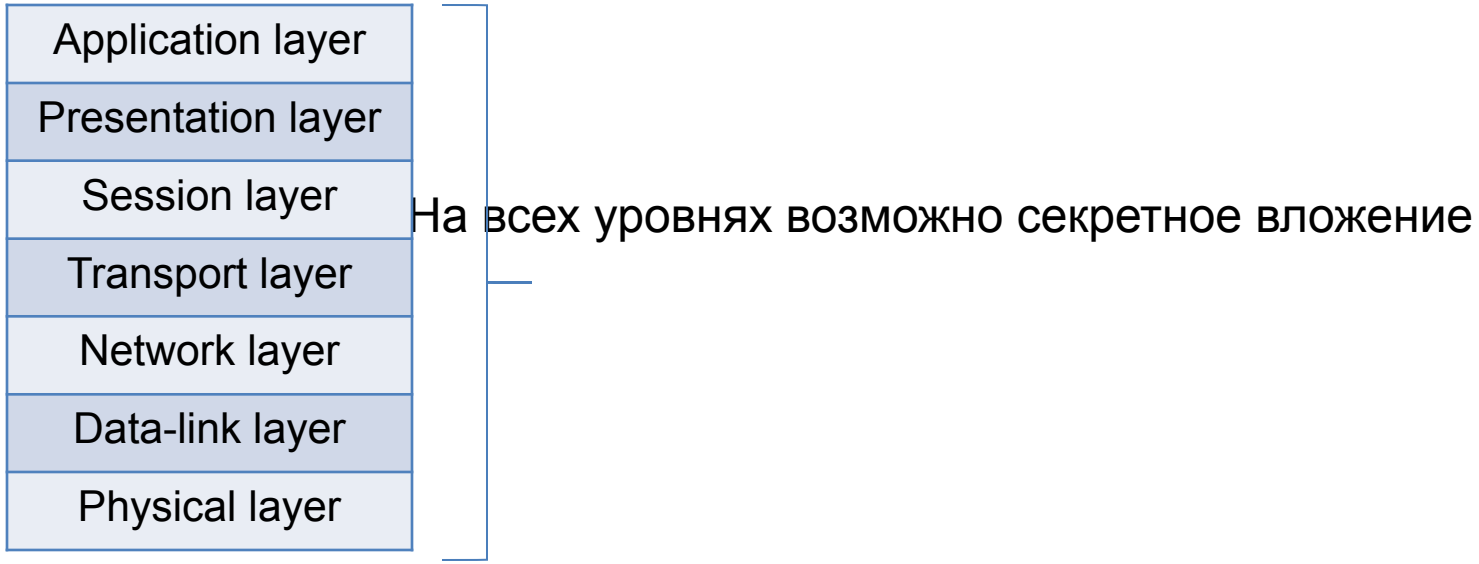
Image №	Up	Down	Left	Right	All	Number of black pixels	Detecting
23	466	653	632	712	2463	798781	Missed
27	439	563	653	631	2286	804339	Missed
41	501	669	753	718	2641	877083	Missed
42	508	659	733	698	2598	851749	Missed
45	795	775	998	1091	3659	946030	Stego - Image
48	508	684	813	710	2715	885834	Missed
50	507	644	759	747	2657	850749	Missed
51	601	719	850	937	3107	810658	Stego - Image
56	505	660	713	709	2587	813280	Missed
57	672	708	890	940	3210	829337	Stego - Image
62	467	560	676	661	2364	754718	Missed
67	499	666	825	791	2781	874827	Missed
73	740	748	992	1079	3559	859889	Stego - Image
75	531	655	800	795	2781	837569	False alarm
76	485	564	758	707	2514	742726	False alarm
79	718	748	1007	1128	3601	895940	Stego - Image
80	591	754	895	823	3063	907919	Missed



- Эксперимент с использованием девяти пар «принтер–сканер» показал, что при отсутствии у стегоаналитика доступа к принтеру и сканеру, с помощью которых были получены покрывающие объекты, можно необнаруженно, вложить порядка 6–8 тыс. бит на страницу текста формата А4.

4.3. Интернет СГС.

Этот тип СГС использует вложения в различные интернет-протоколы, типа TCP/IP.

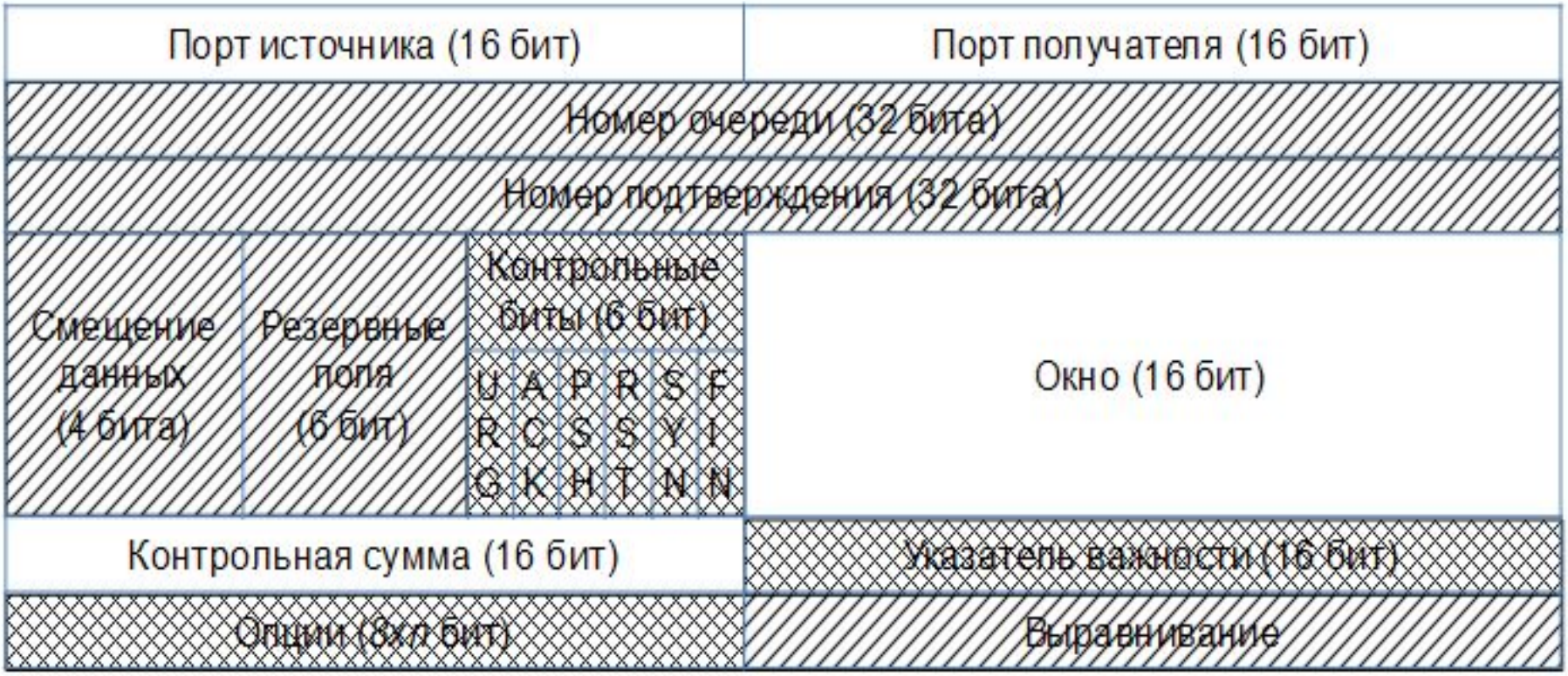


OSI Интернет архитектура

Прикладной	Использование обычных методов стеганографии
Представительный	Погружение данных в поля системных сообщений
Сессионный	Мониторинг чтения пользователями удаленных дисков
Транспортный	Вложение в неиспользованные данные TCP заголовков
Сетевой	Вложение в свободные поля IP пакетов
Уровень данных	Вложение в заголовки фреймов; использование CRC информации
Физический	Конфликтные ситуации с пакетами: “0” – посылка пакета после конфликта с задержкой, “1” – посылка пакета без задержки

Способы вложения скрытой информации на различных уровнях.

Формат TCP заголовка.



Поля, в которых возможно безусловное вложение, заштрихованы, а поля, вложение в которые возможно лишь при определенных условиях, отмечены двойной штриховкой.