УДК 519.21

5.2.2. – Математические, статистические и инструментальные методы в экономике (экономические науки)

РАЗРАБОТКА КОМПЛЕКСНОГО ИНСТРУМЕНТА ДЛЯ ИССЛЕДОВАНИЯ ИНФОРМАЦИОННОЙ СТРУКТУРЫ ТЕКСТОВЫХ И БИНАРНЫХ ДАННЫХ

Мурлина Владислава Анатольевна
Кандидат технических наук, доцент кафедры информационных систем и программирования
РИНЦ-SCIENCE INDEX. SPIN-код: 1256-1843
murlina.v@yandex.ru
Кубанский государственный технологический университет, г. Краснодар, Россия

Щербинин Никита Сергеевич
студент кафедры информационных систем и программирования
nikman11052004@gmail.com
Кубанский государственный технологический университет, г. Краснодар, Россия

Ярыш Александр Борисович
студент кафедры информационных систем и программирования
brobjproyou@gmail.com
Кубанский государственный технологический университет, г. Краснодар, Россия

Червоткин Даниил Вадимович
студент кафедры информационных систем и программирования
chervotkin.dvreserve@gmail.com
Кубанский государственный технологический университет, г. Краснодар, Россия

Паськов Максим Игоревич
студент кафедры информационных систем и программирования
paskov.m.krd@gmail.com
Кубанский государственный технологический университет, г. Краснодар, Россия

Исследованы методы анализа текстовых и бинарных данных, направленные на оценку их информационной структуры, степени случайности и защищённости. Представленный криптоанализатор, основанный на вероятностном подходе, представляет собой комплексный инструмент для исследования защищенности данных и их внутренней организации. Актуальность темы обусловлена стремительным ростом объемов информации и усложнением задач по её анализу в современном мире, где информационные технологии занимают

UDC 519.21

5.2.2. – Mathematical, statistical and instrumental methods in economics (economic sciences)

DEVELOPMENT OF AN INTEGRATED TOOL FOR STUDYING THE INFORMATION STRUCTURE OF TEXT AND BINARY DATA

Murlina Vladislava Anatolievna
Candidate of Technical Sciences, Associate Professor of the Department of Information Systems and Programming
RSCI-SCIENCE INDEX. SPIN code: 1256-1843
murlina.v@yandex.ru
Kuban State Technological University, Krasnodar, Russia

Shcherbinin Nikita Sergeevich
student of the Department of Information Systems and Programming
nikman11052004@gmail.com
Kuban State Technological University, Krasnodar, Russia

Yarysh Aleksandr Borisovich
student of the Department of Information Systems and Programming
brobjproyou@gmail.com
Kuban State Technological University, Krasnodar, Russia

Chervotkin Daniil Vadimovich
student of the Department of Information Systems and Programming
chervotkin.dvreserve@gmail.com
Kuban State Technological University, Krasnodar, Russia

Paskov Maxim Igorevich
student of the Department of Information Systems and Programming
paskov.m.krd@gmail.com
Kuban State Technological University, Krasnodar, Russia

This study focuses on the development of an integrated analytical tool designed to examine the information structure of both textual and binary data. The proposed solution employs probabilistic and statistical methods to evaluate data randomness, internal patterns, and resistance to cryptanalysis. The system is particularly effective for assessing the cryptographic strength of data sequences and identifying structural weaknesses in encryption schemes. One of the core components of this research involves the implementation of advanced randomness testing techniques, including entropy estimation, chi-square tests, and the index of

центральное место. Одним из ключевых аспектов исследования является расчет индекса совпадений, который широко применяется в криптоанализе для различения естественного языка и зашифрованного текста. Этот метод позволяет выявить статистические закономерности, что особенно важно при работе с шифрами, основанными на замене или перестановке символов. Особое внимание уделяется проверке случайности бинарных последовательностей, что имеет решающее значение для оценки качества генерации криптографических ключей и обеспечения надежности алгоритмов шифрования. Эти методы находят широкое применение в современных задачах криптографии, криптоанализа и обеспечения информационной безопасности, способствуя повышению надёжности систем защиты информации и эффективности анализа зашифрованных данных. Реализация данных подходов позволяет не только выявлять уязвимости в существующих системах шифрования, но и разрабатывать рекомендации по их устранению, что особенно важно в условиях растущего числа киберугроз

Ключевые слова: КРИПТОАНАЛИЗ, ЗАЩИТА ДАННЫХ, ЭНТРОПИЯ, ИНДЕКС СОВПАДЕНИЙ, ХИ-КВАДРАТ

coincidence — a critical metric for distinguishing structured linguistic content from encrypted or random data. These methods are essential in modern cryptanalysis, especially when analyzing classical ciphers or evaluating the quality of cryptographic key generation. The relevance of this work stems from the exponential growth of digital data and the increasing complexity of ensuring its confidentiality and integrity. As cyber threats evolve, reliable tools for data analysis and vulnerability detection become indispensable for strengthening information security systems. By combining multiple statistical approaches into a single comprehensive framework, this research contributes to the ongoing development of more efficient cryptanalytic strategies and enhanced data protection mechanisms. The results can be applied in various domains, including cybersecurity, digital forensics, and secure communications

Keywords: CRYPTANALYSIS, DATA SECURITY, ENTROPY, INDEX OF COINCIDENCE, CHI-SQUARE TEST

http://dx.doi.org/10.21515/1990-4665-210-045

Введение. В современном мире информация становится одним из важнейших стратегических ресурсов, играя ключевую роль в развитии общества, экономики и технологий. Однако с увеличением объемов данных и усложнением способов их обработки возрастают и риски, связанные с утечками, несанкционированным доступом и кибератаками. В этих условиях обеспечение безопасности информации становится приоритетной задачей для государственных структур, коммерческих организаций и частных лиц.

Одним из основных направлений защиты данных является разработка и совершенствование методов криптографии, которые позволяют преобразовать информацию в зашифрованный вид, недоступный для злоумышленников. Однако эффективность шифрования напрямую зависит от качества используемых алгоритмов и ключей. Для

оценки надежности таких систем необходимы инструменты анализа, способные выявлять уязвимости и определять степень случайности данных, что является критически важным для генерации криптографических ключей. Поэтому необходимо разработать программное средство, использующее энтропию, индекс совпадений и проверяет случайность бинарных последовательностей для оценки структуры и защищенности данных, которые помогают выявить статистические закономерности в текстовых данных, определить уязвимости в шифровании и оценить качество генерации криптографических ключей, что особенно важно для обеспечения надежности систем защиты информации. Данное средство необходимо специалистам в области информационной безопасности, криптоаналитиков и разработчикам криптографических алгоритмов, а также может быть полезна исследователям, работающим с анализом данных и тестированием алгоритмов шифрования.

Используемые алгоритмы. В рамках исследования разработана программа, которые реализованы ключевые методы криптографического анализа, включая расчет энтропии, вычисление индекса совпадений и проверку случайности бинарных последовательностей.

Вычисление значения энтропии:

Энтропия Шеннона [1] — это мера неопределённости или степени хаотичности информации в сообщении. В теории информации она позволяет оценить минимальный возможный объём данных, необходимый для представления информации без потерь, что играет ключевую роль в сжатии данных.

Для вычисления энтропии текста был реализован метод CalculateEntropyValue, основанный на формуле Шеннона [1]:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \cdot \log_2 P(x_i), \quad (1)$$

где H(X) — энтропия (в битах), $P(x_i)$ — вероятность появления символа x_i в строке, n — количество уникальных символов в строке.

Энтропия измеряется в битах и отражает среднее количество информации, необходимое для определения каждого символа в строке [1]. Высокое значение энтропии свидетельствует о значительной неопределённости, что затрудняет прогнозирование данных и повышает криптографическую стойкость систем. Напротив, низкая энтропия может указывать на избыточность или уязвимости, такие как слабые пароли или предсказуемые шаблоны в данных.

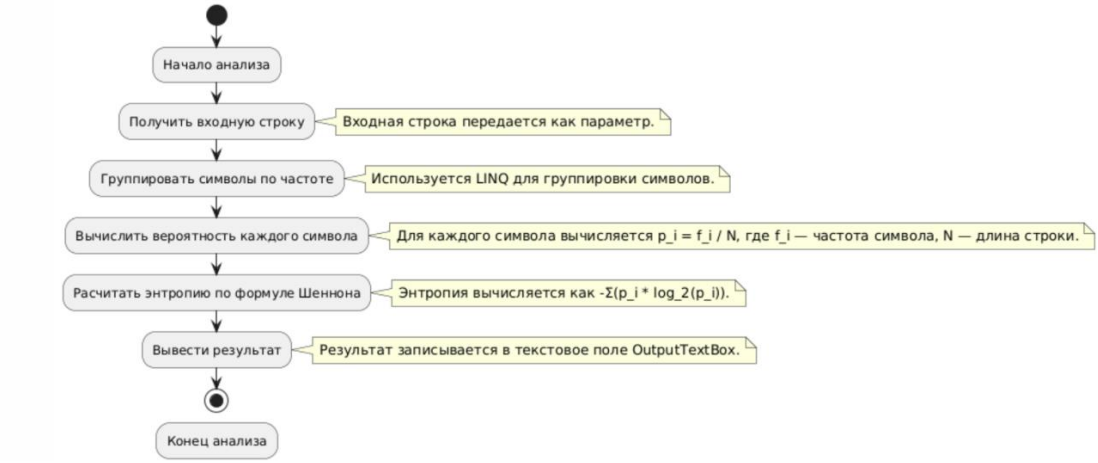

Рисунок 1 – Графическая визуализация CalculateEntropyValue.

Вычисление индекса совпадений:

Для оценки степени случайности и предсказуемости текста используется индекс совпадений. Индекс совпадений (IC) [2] — это статистическая мера, характеризующая вероятность того, что два случайно выбранных символа из текста окажутся одинаковыми. В криптоанализе IC широко применяется для различения естественных языков и зашифрованных сообщений, поскольку в случайном тексте значение индекса обычно ниже, чем в текстах на естественном языке, где частоты символов неравномерны. Таким образом, анализ индекса совпадений помогает оценить степень близости текста к случайному распределению,

что является ключевым при проверке качества шифрования или обнаружении потенциальных уязвимостей в защищённых данных.

Для вычисления индекса совпадений входного текста был реализован метод CalculateIndexOfCoincidence [2]:

$$IC = \frac{\sum_c f(c) \cdot (f(c) - 1)}{N \cdot (N - 1)}, \quad (2)$$

где IC — индекс совпадений, f(c) — частота (количество вхождений) символа, N — общая длина текста.

Если индекс совпадений значительно отличается от ожидаемого значения для случайного текста, это может указывать на наличие структуры или предсказуемости в данных.

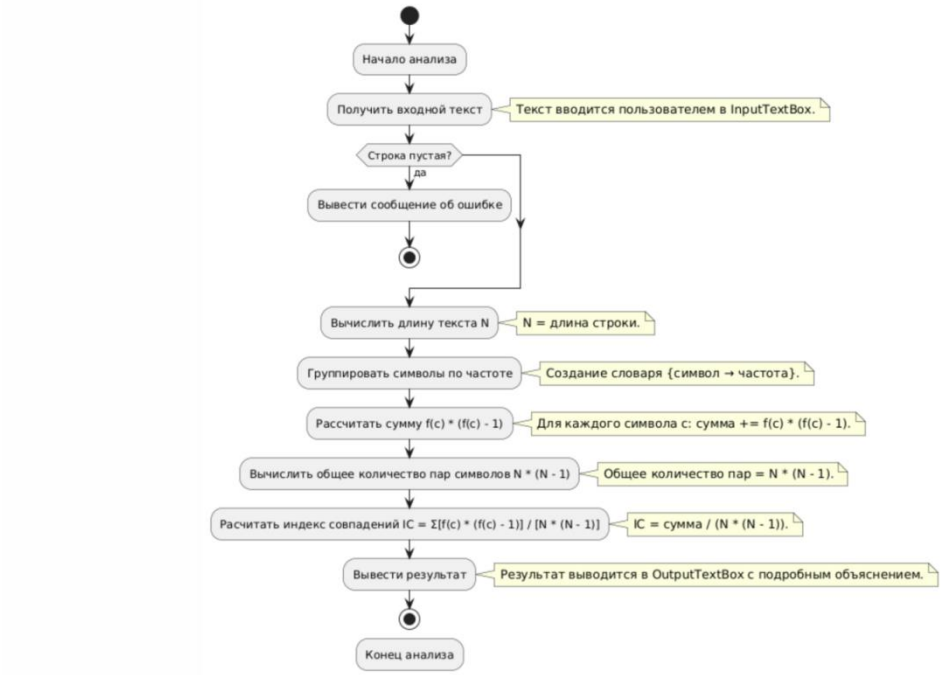

Рисунок 2 – Графическая визуализация CalculateIndexOfCoincidence.

Проверка случайности бинарной последовательности:

Для анализа степени случайности бинарных последовательностей используются специальные методы проверки, позволяющие выявить статистические аномалии и закономерности. Проверка случайности бинарной последовательности [2] — это статистический анализ,

направленный на выявление отклонений от равномерного распределения символов '0' и '1'. Основным инструментом такой проверки является критерий хи-квадрат (χ^2), который сравнивает наблюдаемые частоты битов с ожидаемыми при условии их случайного распределения.

Метод AnalyzeRandomness предназначен для проверки случайности бинарной последовательности с использованием статистического теста хи-квадрат (χ^2) [2]:

$$\chi^2 = \sum_i \frac{(O_i - E_i)}{E_i}, \quad (3)$$

где χ^2 — статистика хи-квадрат, O_i— наблюдаемое количество элементов в категории i, E_i— ожидаемое количество элементов в категории i.

Если полученное значение меньше критического (для выбранного уровня значимости, например 0.05), последовательность считается случайной. В противном случае делается вывод о наличии статистически значимых отклонений от случайности.

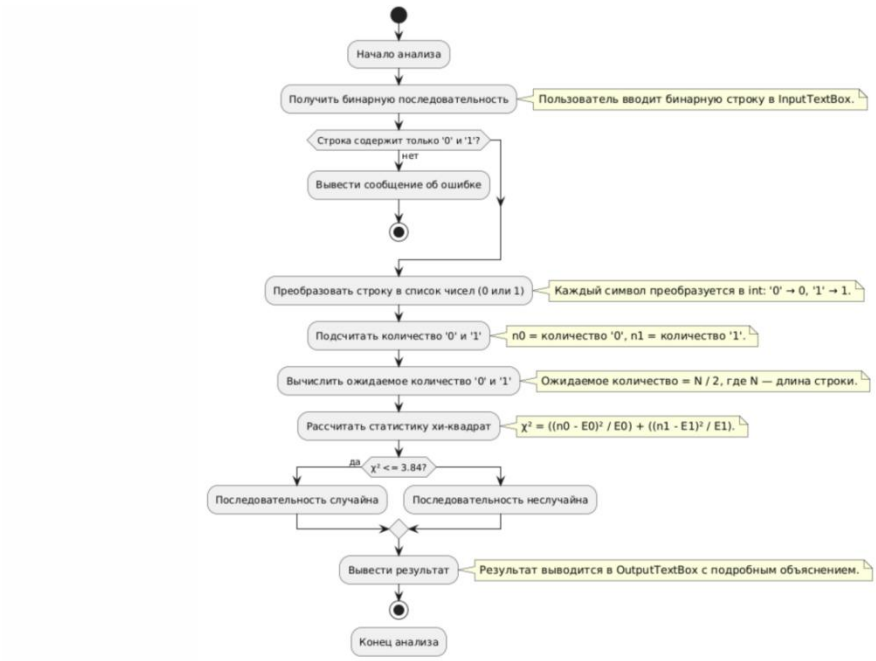

Рисунок 3 – Графическая визуализация AnalyzeRandomness.

Техническое описание. Система криптоанализа представляет собой программное решение, предназначенное для выполнения комплексного анализа текстовых данных с использованием современных методов обработки информации. Она включает в себя набор инструментов для вычисления энтропии, индекса совпадений, проверки случайности бинарных последовательностей. Основные компоненты системы работают синхронно, предоставляя пользователю возможность загружать данные, выбирать необходимый тип анализа и получать детализированные результаты.

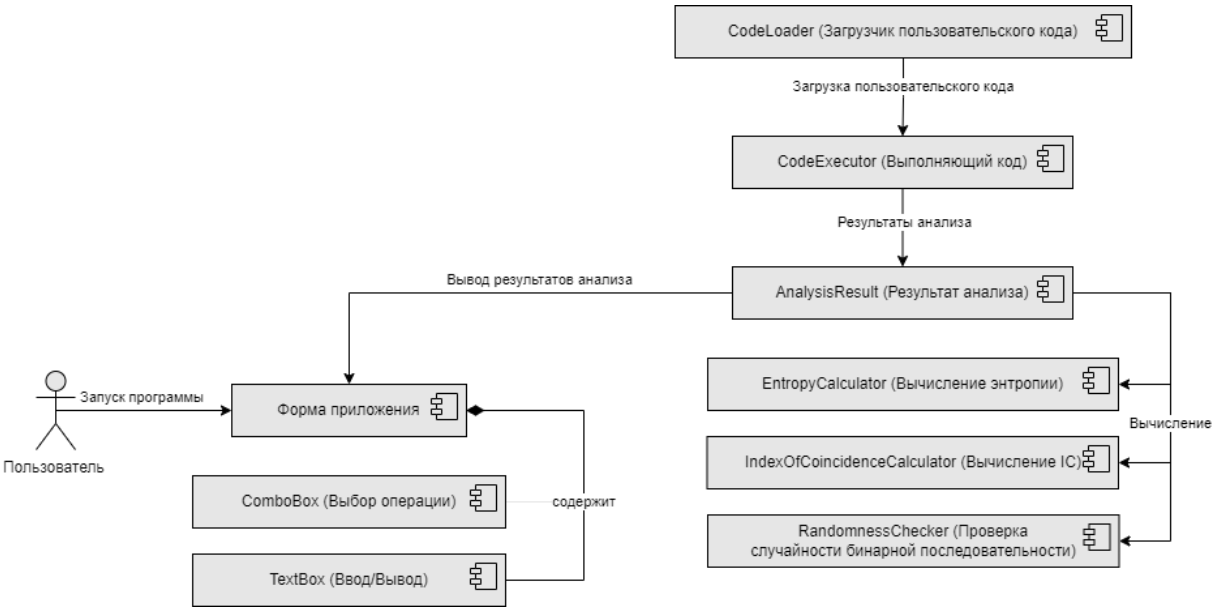

Рисунок 4 – Взаимодействия основных компонентов системы.

В основе системы лежит модульная архитектура, которая позволяет гибко настраивать процесс анализа под конкретные задачи. Пользовательский интерфейс предоставляет доступ к ключевым функциям, таким как расчет энтропии, вычисление индекса совпадений и проверка случайности бинарных последовательностей. Эти операции реализованы с использованием математических моделей, таких как формула Шеннона для энтропии и статистический тест хи-квадрат для проверки случайности. Каждый метод сопровождается подробным

описанием расчетов, что делает систему удобной для использования как профессионалами, так и начинающими исследователями.

Для обеспечения точности и надежности анализа система автоматически проверяет корректность входных данных. При работе с бинарными последовательностями выполняется валидация на предмет наличия только символов "0" и "1". В случае текстовых данных учитываются все уникальные символы, что позволяет проводить анализ даже для сложных алфавитов. Система также учитывает ограничения на длину анализируемых данных, чтобы минимизировать вычислительные затраты и повысить производительность.

Особое внимание уделяется обработке больших объемов данных. Для этого используются оптимизированные алгоритмы, которые позволяют эффективно группировать символы, подсчитывать их частоты и выполнять сложные математические вычисления. При расчете индекса совпадений система анализирует распределение символов в тексте, сравнивая его с теоретическими значениями для случайных и естественных языков. Это позволяет выявить структурированные или зашифрованные данные, что особенно важно в контексте криптографического анализа.

Результаты анализа выводятся в виде подробных отчетов, содержащих численные значения. Система автоматически формирует выводы о степени случайности данных, равномерности распределения символов и других характеристиках.

Результаты анализа данных. Каждый из методов позволил получить объективную характеристику исследуемых данных и сделать выводы об их криптографической устойчивости.

Анализ энтропии строки:

Исходные данные:

Входная строка длиной 68 символов:

eewqkejwqje12hh12h31bdsajwqhkehwkqekhjsadjsajkN!@#J1dnsjadwqehwqjeqw

Результаты:

Энтропия Шеннона: 3,84 бита

Максимально возможная энтропия (для 68 символов): 6,09 бит

В данном случае наблюдается умеренное значение энтропии, что свидетельствует о наличие повторяющихся символов и некоторой структурированности данных. Для достижения максимально возможного значения необходимо равновероятное появление всех символов без повторений.

Расчёт индекса совпадений (IC):

Исходные данные:

Текст длиной 12 символов, в котором символы d, s, a, e, w, q встречаются по 2 раза.

Результаты:

Индекс совпадений (IC): 0,0909

Теоретическое значение для случайного текста: ~0,0385 (для английского алфавита)

Значение индекса совпадений 0,0909 свидетельствует о неравномерном распределении символов и наличии регулярностей, что может быть связано с использованием естественного языка, простых шифров или шаблонных конструкций.

Тест на случайность (χ^2-критерий):

Исходные данные:

Бинарная последовательность длиной 37 бит :

Количество нулей: 19

Количество единиц: 18

Результаты:

Статистика хи-квадрат (χ^2): 0,0270

Критическое значение (p = 0,05, df = 1): 3,8400

Проверка проводилась для гипотезы о равномерном распределении битов в последовательности. Если рассчитанное значение χ^2 меньше критического, последовательность считается случайной. В данном случае условие выполняется: 0,0270 < 3,8400 .

Вывод. В результате проведенного исследования разработана система криптоанализа, позволяющая выполнять комплексный анализ текстовых и бинарных данных с использованием современных методов обработки информации. Реализованные инструменты, такие как расчет энтропии, вычисление индекса совпадений и проверка случайности последовательностей, обеспечивают высокую точность оценки характеристик данных. Применение статистических тестов и алгоритмов анализа делает систему эффективной при работе с различными типами данных, включая малые выборки и зашифрованные последовательности.

Система предоставляет пользователю возможность загружать данные из различных источников, выбирать необходимый тип анализа и получать детализированные отчеты. Результаты выводятся в виде численных значений, графических пояснений и интерпретаций, что позволяет сделать выводы о степени случайности данных, равномерности распределения символов и других ключевых характеристиках. Особое внимание уделено оптимизации вычислений для работы с большими объемами данных, а также обеспечению корректности входных данных.

Предложенное решение может быть успешно применено для анализа защищенности информации, выявления аномалий и поддержки криптографических исследований. Его гибкость и универсальность позволяют использовать систему в широком спектре задач, связанных с защитой данных и оценкой качества шифрования.

ЛИТЕРАТУРА

1. Шавенько, Н. К. Основы теории информации : учебное пособие для студентов высших учебных заведений, обучающихся по специальностям направления подготовки 09.03.02 «Информационные системы и технологии» / Н. К. Шавенько. — Москва : Издательство МИИГАиК, 2019. — 133 с.

2. Шнайер, Б. Прикладная криптография: Протоколы, алгоритмы и исходный код на языке С / Б. Шнайер. — М. : Триумф, 2002. — 816 с.

3. Добробаба, Ю. П. Анализ переходных характеристик системы шестого порядка с кратными корнями характеристического уравнения / Ю. П. Добробаба, А. Г. Мурлин, А. Д. Серкин // Наука. Техника. Технологии (политехнический вестник). – 2019. – № 1. – С. 430-437. – EDN WAMJVI.

4. Павловская Т.А. C#. Программирование на языке высокого уровня. Учебник для вузов. – СПб.: Питер, 2014. – 432 с: ил.

REFERENCES

1. Shaven'ko, N. K. Osnovy teorii informacii : uchebnoe posobie dlja studentov vysshih uchebnyh zavedenij, obuchajushhihsja po special'nostjam napravlenija podgotovki 09.03.02 «Informacionnye sistemy i tehnologii» / N. K. Shaven'ko. — Moskva : Izdatel'stvo MIIGAiK, 2019. — 133 s.

2. Shnajer, B. Prikladnaja kriptografija: Protokoly, algoritmy i ishodnyj kod na jazyke C / B. Shnajer. — M. : Triumf, 2002. — 816 s.

3. Dobrobaba, Ju. P. Analiz perehodnyh harakteristik sistemy shestogo porjadka s kratnymi kornjami harakteristicheskogo uravnenija / Ju. P. Dobrobaba, A. G. Murlin, A. D. Serkin // Nauka. Tehnika. Tehnologii (politehnicheskij vestnik). – 2019. – № 1. – S. 430-437. – EDN WAMJVI.

4. Pavlovskaja T.A. S#. Programmirovanie na jazyke vysokogo urovnja. Uchebnik dlja vuzov. – SPb.: Piter, 2014. – 432 s: il.