УДК 004.89

**Ю.И. БУТЕНКО**

***Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Украина***

# ОБОБЩЕННАЯ МОДЕЛЬ ЯДРА СЕМАНТИЧЕСКОЙ ЦЕЛОСТНОСТИ ЯЗЫКА СТАНДАРТОВ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

*Описана постановка задачи синтеза обобщенной модели ядра семантической целостности для автоматизации процесса обработки текстов стандартов программного обеспечения (ПО) и технической документации к программным продуктам. Применение данной модели в составе компьютеризированной диалоговой системы поддержки принятия решений при сертификации систем с интенсивным использованием программного обеспечения обеспечит повышение эффективности работы сертификационного аудитора за счет снижения доли рутинного труда при формировании нормативного профиля на ПО, а также снижения рисков принятия неверных решений в процессе анализа текстов технической документации на ПО.*

***Ключевые слова**: программное обеспечение, экспертирование ПО, нормативная база, нормативный профиль, синтаксический анализ, семантическая информация, компрессия текста, набор ключевых слов, ядро семантической целостности.*

## Введение

Экспертиза программного обеспечения, являясь основным механизмом оценки его соответствия предъявляемым требованиям и нормативным профилям, в значительной мере определяет реальные возможности в обеспечении необходимого уровня безопасности и качества в отраслях и социальных сферах, связанных с системами критического применения [1, 2].

Процедура оценки ПО проводится на специализированных сертификационных центрах сертификационными аудиторами (СА) [3] и предполагает решение следующих задач:

- формирование нормативного профиля (НП) – гармонизированной с международными и национальными стандартами совокупности требований, предъявляемых к данному проекту или группе проектов. НП могут быть вновь разрабатываемые государственные или отраслевые стандарты, нормативно-методические документы предприятий и общие требования спецификаций ПО;

- реинжиниринг процесса проектирования ПО и его оценка на основе НП;

- статистический анализ исходного текста, заключающийся в определении программных метрик, согласно выбранному нормативному профилю, и выполнении семантического анализа;

- динамический анализ ПО: модульное тестирование методом белого и черного ящиков, и интервальный анализ исполняемого модуля;

- определение степени соответствия исходного кода ПО проектной документации и НП.

В то же время экспертиза ПО является слабо формализованным и слабо структурированным видом профессиональной деятельности СА. Велик субъективизм и влияние опытности СА на результаты итоговых оценок.

Анализ иерархической структуры текстов стандартов и технической документации (ТД) целесообразно осуществлять в два этапа. Первый этап связан с анализом композиционной структуры текста и, в частности, распознаванием нумерации (маркировки) разделов стандартов. Определение нумерованных (маркированных) фрагментов позволяет более точно определить границы предложений и сформировать их в виде, удобном для последующей компьютеризированной обработки, которая дает возможность на втором этапе определять синтаксическую структуру предложения с целью выявления терминологических единиц [2] и их связей для последующего построения соответствующей онтологии предметной области. Очевидно, реализация второго этапа предполагает использование специальной модели представления в единой форме синтаксической структуры обрабатываемых языковых объектов

*Целью статьи* является описание процесса синтеза семантической модели, которая способна отразить в типовой форме (ядра семантической целостности) как фрагменты нормативной базы, так и языковые конструкции текстов технической документации на ПО. В результате последующей машинной обработки полученных результатов может быть сформирован НП требований к объекту сертификации.

## 1. Постановка задачи

Исходными данными для формирования обобщенной модели текстов НБ и ТД для сертификации систем с интенсивным использованием ПО являются представительный набор текстов документов из подмножества профилеобразующей базы, непосредственно относящейся к данной предметной области, а также полный комплект ТД к сертифицируемому программному продукту.

Создаваемая модель должна в автоматическом режиме: декомпозировать анализируемый текст до простых предложений; находить среди членов простых предложений анализируемого текста подлежащие; путем обращения к терминосистеме выделять среди найденных подлежащих ключевые слова и формировать в пределах анализируемого фрагмента текста набор ключевых слов (НКС); формировать «смысловые вехи» (компрессия текста) путем отыскания в анализируемом тексте сказуемых и связанных с ними ключевых слов; путем сравнения соответствующих «смысловых вех» установить наличие ядер семантической целостности в анализируемых текстах НБ и ТД; формировать запрос к СА в случае невозможности определения ядер семантической целостности.

Результатом работы модели являются набор семантической целостности с указанием их позиций в анализируемых текстах.

## 2. Синтаксический анализ нормативной базы программной инженерии и технической документации на ПО

Проведенный синтаксический анализ текстов показал, что синтаксические структуры предложений в текстах стандартов и ТД в большинстве случаев однотипны. Так, например, был проанализирован стандарт «МЭК 60880. Атомные электростанции. Системы контроля и управления, важные для безопасности. Аспекты программного обеспечения компьютерных систем, выполняющих функции категории А; в основных разделах этого стандарта имеется 513 предложений (вводная часть и приложения не были взяты под рассмотрение), из которых 294 являются простыми предложениями,220 – сложными, в том числе к сложным предложениям относим и перечисления, которых в тексте стандарта - 59.

Простые предложения могут быть усложнены причастными и деепричастными оборотами, которые модифицируют значение того члена предложения, к которому они относятся. Сложносочиненные предложения встречаются реже, чем сложноподчиненные. Так как сложносочиненное предложение состоит из 2 и более простых предложений, то их разбор идентичен разбору простых предложений [6].

Сама структура простых предложений в составе сложноподчиненных идентична простому предложению, но следует учитывать особенности подчиненного предложения. В текстах стандартов выявлено три вида придаточных предложений в составе сложноподчиненных, а именно: часть-целое, причина-следствие, условие-причина.

На рис. 1 представлена обобщенная схема всех видов синтаксических конструкций, представленных в текстах стандартов, в которой можно выделить три уровня: уровень сложных предложений, уровень простых предложений и уровень членов предложения.

Алгоритмы синтаксического разбора рассмотрены и систематизированы в [1].

Проведенный синтаксический анализ нормативной базы позволяет определить терминологические единицы, а также связи, существующие между

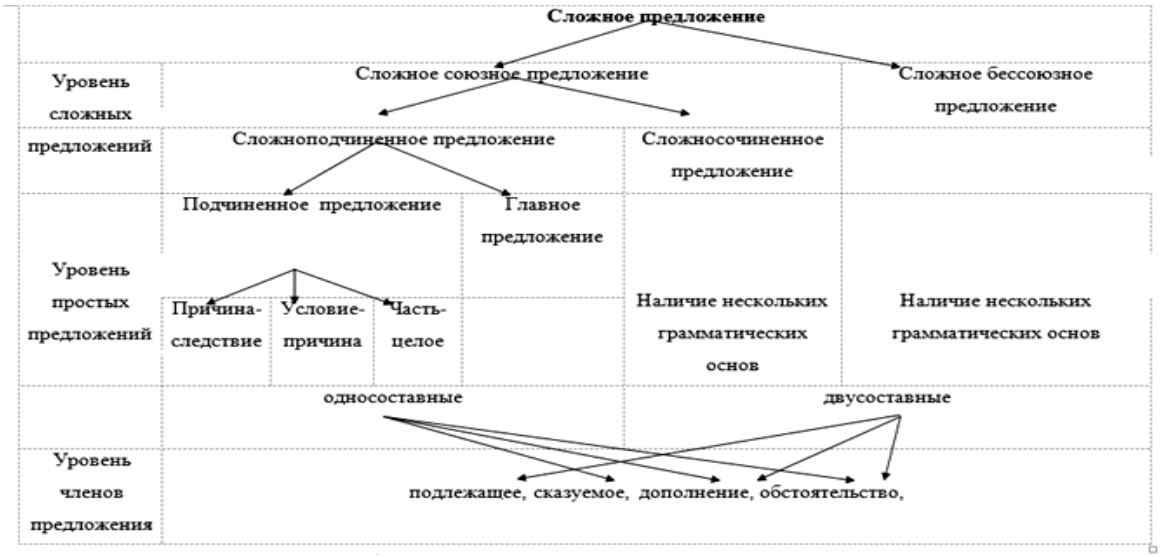


Рис. 1. Виды синтаксических структур в текстах стандартов и ТД на ПО

такими единицами, что в свою очередь создает основу для последующего синтеза машинной процедуры, в которой эти результаты будут использованы для построения онтологической системы.

## 3. Лингвистические основы семантического моделирования языка стандартов и ТД на ПО

Все лексемы как языка стандартов, так и языка на котором написана ТД на ПО, и другие эквивалентные им лексические единицы, в том числе и многословные, делятся прежде всего на два основных типа – предметные и предикатные.

Двум разрядам лексики и, соответственно, двум типам толкований соответствуют и две разные семантические классификации языковых единиц – таксономическая и фундаментальная. Первая предназначена для предметных единиц (например, названий различных объектов живой и неживой природы), а вторая – для предикатных.

Для лексического анализа рассматриваемых языковых объектов используем основные положения, известные в теории фундаментальной классификации. Как известно, фундаментальной называется такая классификация, понятия которой имеют универсальный характер, т.е. используются во всех лингвистических правилах – морфологических (категории вида, залога и наклонения), словообразовательных, синтаксических, семантических, прагматических, сочетаемостных и др. Из них самыми важными являются семантические правила.

Классификация представляет собой нестрогую многоуровневую иерархию с пересечениями классов. Из глагольных категорий на материале русского языка к семантическим различиям между классами оказались чувствительны, кроме вида, залог и наклонение. Кроме того, лексемы, относящиеся к разным классам предикатов, различаются синтаксическими, сочетаемостными свойствами, а также структурами многозначности, словообразовательными типами и типами семантических связей с другими лексемами.

В процессе анализа нормативной базы были выделены следующие **верхние классы**: действия (*подписывать, транспортировать*, проверять, производить,), деятельности (*управлять, руководить, проектировать, разрабатывать*), процессы (*храниться*), положения в пространстве (*размещать*), параметры (*весить*).

К морфологическим свойствам глаголов языка стандартов и языка ТД относятся: изъявительное наклонение, несовершенный вид, действительный или страдательный залог. В большинстве предложений языка стандартов и ТД глаголы используются в настоящем времени.

## 4. Формальное представление обобщенной модели ядра семантической целостности языковых объектов типа НБ и ТД ПО

Обобщенная модель ядра семантической целостности (ЯСЦ) должна обладать достаточной универсальностью, в частности, быть пригодной для сжатия текстовой семантической информации, являясь основой построения моделей для экспертной системы поддержки принятия решений сертификационным аудитором при экспертировании ПО.

Начальный этап создания модели ЯСЦ состоит в выборе множества базовых категорий. К базовым в данном случае необходимо отнести следующие категории: «предмет»; «свойства предмета»; «отношение предмета к другим предметам».

*Анализ базовой категории «предмет».* Представим базовую категорию «предмет» в виде множества

$$Q = \{Q_1, Q_2, ..., Q_k\},$$

где $Q_1$ – основной предмет;

$Q_2$, ..., $Q_k$ – вспомогательные (взаимодействующие) предметы.

Представление предмета в виде множества предметов позволяет формулировать сложные запросы одновременно в одном запросе и выражает потребность пользователя в двух, трех и т. д. равных по значимости предметах.

В общем случае наличие нескольких предметов в модели ЯСЦ позволяет формально осуществить поиск по любому из них или сочетанию нескольких, не отдавая предпочтения тем или иным предметам.

*Анализ базовой категории «свойства предмета».*

Обозначим множество свойств предмета как $J$. Тогда

$$J = \{J_1, J_2\},$$

где $J_1$, $J_2$ - качественная и количественная определенности соответственно.

Представим качественную определенность предмета в виде множества

$$J = \{N_1\ N_2, N_3, N_4\},$$

где $N_1$ – функциональное назначение;

$N_2$ – область применения;

$N_3$ – отличительный признак;

$N_4$ – принцип действия (взаимодействие элементов, обеспечивающих выполнение заданной функции).

Совокупность свойств, указывающих на размеры предмета, на его величину и др., составляет его количество.

Таким образом, количественную определенность предмета можно характеризовать его структурой и параметрами. Тогда

$$J_2 = \{F_1, F_2\},$$

где $F_1$ – структура предмета; $F_2$ – параметры предмета.

Естественно, что предмет имеет множество параметров $F_2 = \{F_{21}, F_{22}, ...\}$.

Качество и количество неразрывно взаимосвязано между собой и образуют меру, как определенный диапазон, в котором тот или иной параметр модели ЯСЦ имеет допустимое значение.

Для обеспечения гибкости модели ЯСЦ необходимо включить в ее состав обобщенный аспект дополнительных сведений. С его помощью можно отразить запросы по другим свойствам предмета.

Таким образом, категорию свойства каждого предмета, представленного в модели ЯСЦ можно записать в виде

$$J = \{N_1, N_2, N_3, N_4, F_1, F_2, D\},$$

где D - дополнительные сведения.

*Анализ базовой категории «отношение предмета к другим предметам».* Как известно, отношение - категория, характеризующая взаимозависимость элементов определенной системы; оно имеет объективный и универсальный характер.

Формальное описание отношений в естественных и искусственных языках для повествовательной формы выражений принято осуществлять предикатами. В обобщенном виде, предикат от n переменных (от n неопределенных терминов или слов) выражают формулой

$$P(x_1, x_2, ..., x_n), \quad n \geq 0.$$

При n=0 предикат совпадает с высказыванием; при n =1 предикат представляет собой свойство в узком смысле (одноместный предикат); при n=2 – свойство пары (двухместный предикат или бинарное отношение); при n =3 - свойство тройки (трехместный предикат или тернарное отношение) и т.д. Выражения «x – оператор программы», «x принадлежит y»; «x – часть y и z» служат соответственно примерами одно-, двух- и трехместного предикатов.

Формальной основой модели является матрица, по вертикали которой содержатся аспекты (категории), количественно отображающие посредством знаков полноту представления семантической информации (в дальнейшем просто полнота); по горизонтали – позиции, количественно отображающие посредством знаков точность представления семантической информации (в дальнейшем просто точность).

Аспект $\alpha$характеризует определенное свойство объекта и не поддается дальнейшему смысловому делению. В математической интерпретации аспект представляет собой кортеж знаков (букв, слов, символов и др.), длина которого может быть произвольной. Так, кортежем длины n является запись вида

$$\alpha = < a_1, a_2, ..., a_n >,$$

где $a_1$, $a_n$ – первая и последняя компоненты соответственно.

Применительно к текстовой форме семантической информации аспекты представляются кортежами знаков типа букв, цифр, символов и т. д. из различных алфавитов: русского, латинского, греческого, специального.

Свойство аспекта быть кортежем подтверждается следующими его свойствами:

$$\alpha = \{a_i \in \alpha : a_i, \rightarrow R(a_i)\},$$

где $R(a_i)$ - отношение «быть упорядоченным по местам». При этом $\min \alpha = a_1$; $\max\alpha = a_n$; $i = \overline{1, n}$.

$$\forall a_i (a_i \in \alpha)\{Q(a_i) \vee \neg Q(a_i)\},$$

где $Q(a_i)$ – отношение «быть одинаковыми».

Действительно, в названии аспекта или его значении знаки (буквы, цифры и др.) могут быть одинаковыми и разными.

В информационном плане аспект является элементом слова С.

$$\forall a_i : a_i \in \alpha \rightarrow a_i \in C.$$

Каждый аспект характеризуется точностью, т. е. содержит определенное число знаков.

Полнота $C_в$ представляет собой объединение кортежей знаков, выражающих слово. Слово С характеризует предмет, его свойства и отношения. В семантическом плане слово состоит из аспектов. Глубина характеристики объекта определяется количеством аспектов в слове, которое оценивается объемом сведений, необходимым для описания объекта в рамках решаемой задачи. В общем случае число аспектов определяется на основании анализа статистических данных. В частном случае возможно строгое аналитическое определение числа аспектов.

В математической интерпретации слово представляет собой в общем случае кортеж знаков, длина которого может быть произвольной. Выражение вида $C = < \alpha_1, \alpha_2, ..., \alpha_m >$, являясь словом, представляет собой кортеж длины *m*.

Описанные выше элементы модели ЯСЦ транслируются в соответствующие элементы системы онтологий, а именно: в онтологию верхнего уровня для описания нормативной базы сертификации, в онтологию предметных областей для представления знаний каждого отдельного нормативного документа, в онтологию источника знаний, который описывает терминосистему предметной области сертификации, в онтологию задач и методов. При

этом в онтологии задач и методов представлен метод извлечения знаний из нормативных документов, в основу которого положена валентность глаголов. Онтология запроса описывает конкретный запрос пользователя к объекту сертификации а, онтология-приложение – срез онтологии предметной области и онтологии задач и методов к онтологии запроса. Онтологическая система, построенная на основе модели ЯСЦ, является ядром диалоговой системы поддержки принятия решений аудитором сертификационного центра.

## Выводы

Описанная в статье модель ядра семантической целостности языковых объектов из предметной области «Сертификация систем с интенсивным использованием ПО» является основой для создания системы онтологий, содержащих знания концептуального характера о смысловой структуре нормативной базы и технической документации на системы, подвергаемые процедуре сертификации. Применение модели ЯСЦ обеспечивает возможность реализации онтологического среза и формирования на его основе отчетов в ответ на запросы пользователя, которым является сертификационный аудитор.

## Литература

*1. Ахо, А. Теория синтаксического анализа, перевода и компиляции [Текст]: пер. с англ. / А. Ахо, Дж. Ульман. – М.: Изд-во «Мир», 1978. – 616 с.*

*2. Даниленко, В.П. Лексические требования к стандартизуемой терминологии [Текст] / В.П. Даниленко // Терминология и норма. – М.: Наука, 1972. – С. 5-32.*

*3. Конорев, Б.М. Концепция и принципы реализации интегрированной инструментальной системы для поддержки экспертизы и независимой верификации критического программного обеспечения [Текст] / Б.М. Конорев, В.С. Харченко, Г.Н. Чертков // Государственный комитет ядерного регулирования Украины, Государственный центр регулирования качества поставок и услуг, Сертификационный центр АСУ, Харьков, 2003. – 60 с.*

*Поступила в редакцию 03.06.2013, рассмотрена на редколлегии 12.06.2013*

**Рецензент:** д-р техн. наук, проф., проф. кафедры "Программная инженерия" С.Ю. Шабанов-Кушнаренко, ХНУРЭ, Харьков.

### УЗАГАЛЬНЕНА МОДЕЛЬ ЯДРА СЕМАНТИЧНОЇ ЦІЛІСНОСТІ МОВИ СТАНДАРТІВ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

***Ю.І. Бутенко***

Описана постановка задачі синтезу узагальненої моделі ядра семантичної цілісності для автоматизації процесу обробки текстів стандартів програмного забезпечення (ПО) і технічної документації до програмних продуктів. Застосування даної моделі в складі комп'ютеризованої діалогової системи підтримки прийняття рішень при сертифікації систем з інтенсивним використанням програмного забезпечення забезпечить підвищення ефективності роботи сертифікаційного аудитора за рахунок зниження частки рутинної праці при формуванні нормативного профілю на ПЗ, а також зниження ризиків прийняття невірних рішень у процесі аналізу текстів технічної документації на ПЗ.

**Ключові слова**: програмне забезпечення, експертування ПО, нормативна база, нормативний профіль, синтаксичний аналіз, семантична інформація, компресія тексту, набір ключових слів, ядро семантичної цілісності.

### THE GENERALIZED MODEL OF THE KERNEL SEMANTIC INTEGRITY OF LANGUAGE OF STANDARDS OF THE SOFTWARE

***Yu.I. Butenko***

The problem definition of synthesis of the generalized model of a kernel semantic integrity for automation of process of text manipulation of software standards and technical documentation to software products is described. Application of this model as a part of the computerized dialogue decision making support system in case of certification of systems with an intensive use of the software will provide increase of overall performance of the certified auditor at the expense of lowering of a share of routine work when forming a normative profile on a software, and also lowerings of risks of acceptance of incorrect decisions in the course of the analysis of texts of technical documentation on a software.

**Key words**: software, experts the software's, normative base, normative profile, parsing, semantic information, text compression, set of keywords, kernel of semantic integrity.

**Бутенко Юлия Ивановна** – аспирант каф. Инженерии программного обеспечения, Национального аэрокосмического университета им. Н.Е. Жуковского «Харьковский авиационный институт», Харьков, Украина; e-mail: iuliiabutenko@yandex.ru