УДК 004.89

Семантическая модель языковых объектов для автоматизации процесса сертификации систем критического применения

© Ю.И. Бутенко, И.В. Шостак

Национальный аэрокосмический университет
им. Н.Е. Жуковского «ХАИ», Харьков, 61070, Украина

Представлены результаты исследования 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 ПО обеспечит повышение эффективности работы сертификационного аудитора в результате сокращения доли рутинного труда при формировании нормативного профиля на ПО, а также снижения рисков принятия неверных решений в процессе анализа текстов технической документации на ПО. Приведены лингвистические основы семантического моделирования языка стандартов и технической документации на ПО. Дано формальное представление обобщенной модели ядра семантической целостности языковых объектов типа «Нормативная база и техническая документация 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. Одной из особенностей современного этапа развития техники является распространение систем критического применения, сбои и отказы в работе которых несут потенциальную угрозу природе и человеческому обществу [1]. При этом ключевая роль в обеспечении безопасности принадлежит информационно-управляющим системам (ИУС), выполняющим функции предотвращения, защиты и ликвидации последствий аварийных ситуаций [2].

По результатам исследований, каждый пятый отказ оборудования атомной электростанции связан с неисправностями оборудования ИУС так же, как каждая пятая авария ракетно-космической техники обусловлена неисправностями компьютерной системы управления. Шесть из семи отказов этих систем, которые привели к авариям ракетно-космических комплексов, вызваны дефектами программных средств [3]. Особую важность для ИУС комплексов критического применения представляет сертификация ПО, которая выполняется в целях проверки соответствия программного продукта и процессов

его разработки требованиям международных и национальных нормативных документов [4].

Процедура оценки ПО проводится в специализированных сертификационных центрах сертификационными аудиторами (СА) [5]. Она предполагает решение следующих задач:

- формирование нормативного профиля (НП) — гармонизированной с международными и национальными стандартами совокупности требований, предъявляемых к данному проекту или группе проектов. Это 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 НП 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 экспертизу ПО можно считать слабо формализованным и слабо структурированным видом профессиональной деятельности СА. Велики роль субъективизма и влияние опытности СА на итоговые оценки. При этом наиболее критичной по отношению к конечному результату в деятельности СА является процедура анализа базы нормативных документов в целях формирования НП, непосредственно относящегося к объекту сертификации.

Анализ иерархической структуры текстов стандартов и технической документации (ТД) целесообразно осуществлять в два этапа [6]. Первый этап связан с анализом композиционной структуры текста, в частности с распознаванием нумерации (маркировки) разделов стандартов. Определение нумерованных (маркированных) фрагментов позволяет более точно установить границы предложений и сформировать их в виде, удобном для последующей компьютеризированной обработки, которая дает возможность на втором этапе определять синтаксическую структуру предложения в целях выявления терминологических единиц [7, 8] 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 работы является описание процесса синтеза семантической модели, которая позволяет отразить в типовой форме (ядра се-

мантической целостности) как фрагменты нормативной базы (НБ), так и языковые конструкции текстов ТД на ПО. В результате последующей машинной обработки полученных результатов может быть сформирован НП требований к объекту сертификации.

Постановка задачи. Исходными данными для формирования обобщенной модели текстов НБ и ТД для сертификации систем с интенсивным использованием ПО служат 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 (ЯСЦ) в анализируемых текстах НБ и ТД; формировать запрос к СА в случае невозможности определения ЯСЦ.

Результатом работы модели является набор ЯСЦ с указанием их позиций в анализируемых текстах.

Синтаксический анализ НБ программной инженерии и ТД на ПО. Проведенный синтаксический анализ текстов показал, что синтаксические структуры предложений в текстах стандартов и ТД в большинстве случаев однотипны. Так, 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 рассматривались), из которых 294 являются простыми предложениями, 220 — сложными, в том числе к сложным предложениям были отнесены и перечисления, которых в тексте стандарта было 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. Поскольку сложносочиненное предложение состоит из двух простых предложений и более, то их разбор идентичен разбору простых предложений [9].

По структуре простые предложения в составе сложноподчиненных идентичны 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 рисунке приведена обобщенная схема всех видов синтаксических конструкций, представленных в текстах стандартов. В схеме можно выделить три уровня: уровень сложных предложений, уровень простых предложений и уровень членов предложения.

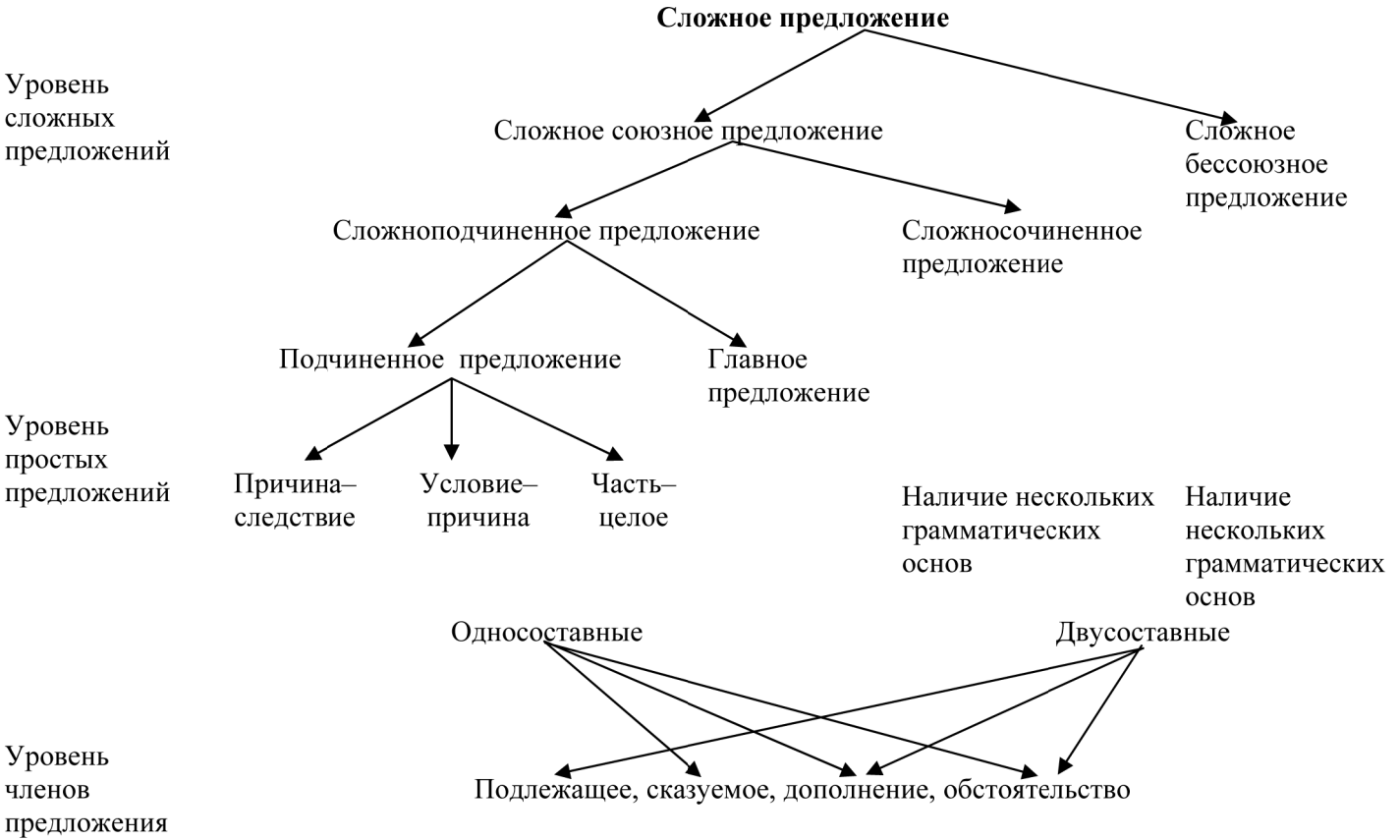
Алгоритмы синтаксического разбора рассмотрены и систематизированы в [10].

Проведенный синтаксический анализ НБ позволяет определить терминологические единицы, а также связи, существующие между ними, что, в свою очередь, создает основу для последующего синтеза машинной процедуры, в которой эти результаты будут использованы для построения онтологической системы.

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, подразделяют на два основных типа — предметные и предикатные. Двум разрядам лексики (двум типам толкований) соответствуют две разные семантические классификации языковых единиц — таксономическая и фундаментальная. Первая предназначена для предметных единиц (например, названий объектов живой и неживой природы), а вторая — для предикатных [11].

Для лексического анализа рассматриваемых языковых объектов используем основные положения теории фундаментальной классификации. Как известно, фундаментальной называют такую классификацию, понятия которой имеют универсальный характер, т. е. применяются 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 (помимо вида) залог и наклонение. Кроме того, лексемы, относящиеся к разным классам предикатов, различаются синтаксическими и сочетаемостными свойствами, а также структурами многозначности, словообразовательными типами и типами семантических связей с другими лексемами [12].



Виды синтаксических структур в текстах стандартов и ТД на ПО

В процессе анализа НБ 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 относят: изъявительное наклонение, несовершенный вид, действительный или страдательный залог. В большинстве предложений языка стандартов и ТД глаголы используются в настоящем времени.

Формальное представление обобщенной модели ЯСЦ языковых объектов типа НБ и ТД ПО. Обобщенная модель ЯСЦ должна обладать достаточной универсальностью, в частности быть пригодной для сжатия текстовой семантической информации, являясь основой построения моделей для экспертной системы поддержки принятия решений СА 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 по сочетанию нескольких, не отдавая предпочтения тем или иным предметам.

Анализ базовой категории «свойства предмета». Обозначим через J множество свойств предмета. Тогда

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

где J_1, J_2 — качественная и количественная определенности.

Представим качественную определенность предмета в виде множества, состоящего из четырех элементов:

$$J_1 = \{N_1, N_2, N_3, N_4\},$$

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 и другие параметры, составляет его количество.

Таким образом, количественную определенность предмета можно характеризовать его структурой и параметрами. Тогда количественная определенность представим в виде множества с двумя элементами:

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

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 связаны 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) представляется в виде зависимости

$$P(x_0, x_1, x_2, ..., x_n), \quad n \geq 0.$$

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, в которой по вертикали расположены аспекты (категории), количественно отображающие посредством знаков полноту представления семантической информации (далее — полнота), по горизонтали — позиции, количественно отображающие посредством знаков точность представления семантической информации (далее — точность).

Введем для дальнейших рассуждений понятие «аспект» α, характеризующий определенное свойство объекта и не поддающийся 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 из различных алфавитов: русского, латинского, греческого, специального.

Свойство аспекта быть кортежем подтверждается следующим его свойством:

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

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

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

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 C:

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

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

Слово *C* 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 в общем случае представляет собой 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 [13–15], а именно: в онтологию верхнего уровня для описания НБ 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 СА.

Выводы. Описанная модель ЯСЦ языковых объектов из предметной области «Сертификация систем с интенсивным использованием ПО» является формальной основой для создания системы онтологий, содержащих знания концептуального характера о смысловой структуре НБ и ТД на системы, подвергаемые процедуре сертификации. Применение модели ЯСЦ позволит обеспечивать возможность реализации онтологического среза и формирования на его основе отчетов в ответ на запросы пользователя, которым является СА.

ЛИТЕРАТУРА

[1] Харченко В.С., Ястребинецкий М.А., Васильченко В.Н. Нормирование и оценка безопасности информационных и управляющих АЭС: регулирующие требования к программному обеспечению. *Ядерная и радиационная безопасность*, 2002, № 1, с. 18—33.

[2] Конорев Б.М., Сергиенко В.В., Чертков Г.Н., Алексеев Ю.Г. Доказательная независимая верификация и оценка скрытых дефектов критического программного обеспечения на основе диверсифицированного измерения инвариантов. *Радиоэлектронные и компьютерные системы*, 2009, № 7, с. 192—199.

[3] Харченко В.С., Скляр В.В., Тарасюк О.М. Анализ рисков аварий для ракетно-космической техники: эволюция причин и тенденций. *Радиоэлектронные и компьютерные системы*, 2003, № 3, с. 135—149.

[4] Тарасюк О.М. *Методы и инструментальные средства метрико-вероятностной оценки качества программного обеспечения информационно-управляющих систем критического применения.* Дис. … канд. техн. наук. Харьков, 2004, 204 с.

[5] Шостак И.В., Шостак И.В., Бутенко Ю.И., Шостак Е.И. Знание-ориентированные методы формирования нормативных профилей к системам критического применения на основе онтологий. *Радиоэлектронные и компьютерные системы*, 2010, № 5, с. 104—108.

[6] Шостак И.В., Бутенко Ю.И. Подход к автоматизации процесса формирования нормативного профиля при сертификации программных продуктов. *Системы обработки информации*, 2010, № 8 (89), с. 122—126.

[7] Даниленко В.П. *Лексические требования к стандартизуемой терминологии*. *Терминология и норма.* Москва, Наука, 1972, с. 5—32.

[8] Даниленко В.П. *Русская терминология. Опыт лингвистического описания*. Москва, Наука, 1977, 246 с.

[9] Нелюбин Л.Л. Перевод и прикладная лингвистика. Москва, Высш. шк., 1983, 207 с.

[10] Ахо А., Ульман Дж. Теория синтаксического анализа, перевода и компиляции. Москва, Мир, 1978, 616 с.

[11] Апресян Ю.Д. *Исследования по семантике и лексикографии*. *Т. 1: Парадигматика*. Москва, Языки славянских культур, 2009, 568 с.

[12] Апресян Ю.Д. *Избранные труды*. *В 2 т. Т. 1. Лексическая семантика (синонимические средства языка)*. Москва, Восточная литература, 1995, 472 с.

[13] Лукашевич Н.В. Тезаурусы в задачах информационного поиска. Москва, Изд-во Моск. ун-та, 2011, 512 с.

[14] Левашова Т.В., Пашкин М.П., Смирнов А.В. Управление онтологиями (базами знаний). Ч. I. *Известия РАН. Сер. Теория и системы управления.* 2003, № 4, с. 132—146.

[15] Левашова Т.В., Пашкин М.П., Смирнов А.В. и др. Управление онтологиями. Ч. II. *Известия РАН. Сер. Теория и системы управления,* 2003, № 5, с. 89—101.

Статья поступила в редакцию 05.07.2013

Ссылку на эту статью просим оформлять следующим образом:
Бутенко Ю.И., Шостак И.В. Семантическая модель языковых объектов для автоматизации процесса сертификации систем критического применения. *Инженерный журнал: наука и инновации*, 2013, вып. 12. URL: http://engjournal.ru/catalog/appmath/hidden/1165.html

Бутенко Юлия Ивановна родилась в 1987 г., окончила Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ» в 2009 г. Аспирантка кафедры «Инженерия программного обеспечения» Национального аэрокосмического университета им. Н.Е. Жуковского «ХАИ». Автор 10 работ в области искусственного интеллекта и прикладной лингвистики. e-mail: iuliiabutenko@yandex.ru

Шостак Игорь Владимирович родился в 1961 г., окончил Харьковский институт радиоэлектроники в 1983 г. Д-р техн. наук, профессор кафедры «Инженерия программного обеспечения» Национального аэрокосмического университета им. Н.Е. Жуковского «ХАИ». Автор около 150 научных работ в области искусственного интеллекта и информационных технологий. e-mail: iv_shostak@rambler.ru