АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322.4’36

А.В. Кан, В.Д. Ревина, В.И. Руснак, Ал-др А. Хорошилов, А.А. Хорошилов

Автоматическое формирование синтаксической модели языка для задач машинного перевода и информационного поиска*

Описывается формальная модель синтаксической структуры текстов на основе обобщенных синтагм и методы её автоматического построения. Предложено новое решение ряда актуальных задач автоматической обработки текстовой информации в области информационного поиска и машинного перевода, а также подход к проблеме унификации синтаксических структур предложений и синтаксических конструкций, входящих в состав предложений.

Ключевые слова: *флективные классы слов, машинная грамматика, морфологический анализ, семантико-синтаксический анализ, формальная модель синтаксической структуры текста, обобщенные синтагмы*

ВВЕДЕНИЕ

Современное общество все чаще называют информационным и для этого есть серьезные основания: в последние десятилетия значительно увеличились объемы разноязычной информации, используемой во всех сферах человеческой деятельности, и появились мощные технические средства ее передачи и обработки [1]. Но в области смысловой обработки информации успехи значительно скромнее и зависят они, прежде всего, от достижений в изучении человеческого мышления, процессов речевого общения между людьми и от умения моделировать эти процессы на ЭВМ. Успешному развитию методов смысловой обработки текстовой информации мешают неправильные представления о природе информации и о смысловой структуре текстов. До сих пор бытует неправильная точка зрения на средства обозначения наименований понятий в текстах: считается, что основные средства их обозначения – это отдельные слова, а не устойчивые фразеологические словосочетания. Такой взгляд на смысловую структуру текстов в XX в. задержал развитие систем машинного перевода текстов с одних языков на другие, как минимум, на несколько десятилетий, и даже в настоящее время некоторые разработчики этих систем продолжают считать, что при переводе следует ориентироваться на “значения слов”.

Как показали исследования, во многих языках мира (например, английском, испанском, немецком, русском, французском) количество различных наименований понятий достигает нескольких сотен миллионов. Большинство из них обозначается фразеологическими словосочетаниями, суть которых не сводится к смыслу составляющих их слов. Слова, входящие в состав словосочетаний, обозначают лишь некоторые признаки понятий, позволяющие отличать их друг от друга, но не исчерпывающие их содержания. Содержание понятий в полном объёме может быть интерпретировано только при условии, что «всё связано со всем».

При создании систем автоматической обработки текстовой информации очень важно исходить из правильных представлений о смысловой структуре языка и речи. По современным представлениям наиболее устойчивыми единицами смысла являются устойчивые фразеологические словосочетания, занимающие центральное место в языке и речи и являющиеся теми базовыми строительными блоками, на основе которых формируются смысловые единицы более высоких уровней. Второй по значимости единицей смысла является предложение. Из предложений формируются различного рода сверхфразовые единства, которые представляются в виде последовательностей предложений – связного текста.

Основная черта предложений – это их предикативность, т. е. свойство утверждения наличия у объектов определенных признаков и их отношений. Свойством предикативности обладают и высказыва-

* Работа выполнена при поддержке гранта РФФИ (проект № 18-37-00110 мол_а)

ния, формулируемые на формализованных языках. Это позволяет сделать вывод, что в основе и предложений на естественном языке, и формализованных логических высказываний лежит предикатно-актантная структура, компонентами которой являются понятия-предикаты (признаки и отношения) и понятия-актанты, выступающие в роли описываемых объектов. В естественных и в формализованных языках предикатно-актантные структуры – это те смысловые инварианты, которые позволяют осуществлять автоматический перевод текстов с естественных языков на формализованные и с формализованных на естественные. Они также позволяют осуществлять автоматический перевод текстов с одних естественных языков на другие.

Смысл текста выражается с помощью единиц смысла, входящих в его состав. В естественных языках базовыми единицами смысла являются понятия. Поэтому центральной процедурой любых систем автоматической смысловой обработки текстов должен быть их семантико-синтаксический концептуальный анализ, реализованный как фразеологический концептуальный анализ на основе мощных словарей наименований понятий. При этом следует опираться на адекватные семантико-синтаксические модели текстов, в которых понятия представляются преимущественно фразеологическими словосочетаниями (например, синтаксическая модель “дерево зависимостей” для этой цели непригодна, так как в ней текст расчленяется на отдельные слова, в результате чего его понятийная структура разрушается).

ОБЗОР СИНТАКСИЧЕСКИХ МОДЕЛЕЙ ЯЗЫКА

Несмотря на значительное число разработанных синтаксических моделей текстов, все они являются условным отражением их структуры и ориентированы на решение конкретных практических задач автоматической обработки. Поэтому любая синтаксическая модель всегда неполна и содержит ошибки, связанные с невозможностью установления однозначных правил синтаксического разбора текстов. При этом часто при построении реальных процедур синтаксического анализа используются элементы различных моделей (например, модели дерева зависимостей и модели членов предложения).

Сегодня существует много моделей языка, используемых в системах обработки текстов на естественных языках, при построении которых широко используются методы математической статистики. Для естественного языка одной из базовых является статистическая модель, основанная на n-граммах (n-граммы представляют собой последовательность из n слов). Эта модель языка используется для предсказания элемента в последовательности, содержащей n-1 предшественников. Вероятность появления n-граммы в тексте можно оценить, зная частоту ее встречаемости в обучающем массиве [2]. Идеи, заложенные при разработке данной модели, лежат в основе современных моделей языка, основанных на статистическом анализе текста. Подробный обзор этих моделей приведен в работе [3], где анализируется несколько вариантов статистических моделей синтаксической структуры текстов.

Модели, основанные на классах, – в них используется функция, которая отображает каждое слово на класс. В этом случае оценка условной вероятности может быть аппроксимирована по n-грамме класса. Функция отображения слова на класс может быть определена вручную с использованием некоторой морфологической информации или же с помощью методов, позволяющих определить функцию отображения автоматически по корпусу текстов [4].

Триггерные модели рассматривают взаимоотношение пар слов в более длинном контексте. В этом случае наличие инициирующего слова увеличивает вероятность появления другого слова (называемого целевым), с которым оно связано. Упрощенной версией триггерных пар является кэш-модель, которая увеличивает вероятность появления слова в соответствии с тем, как часто это слово использовалось в тексте, поскольку считается, что, употребив конкретное слово, автор будет применять это слово еще раз либо потому, что оно характерно для конкретной темы, либо потому, что это слово является частью его лексикона. Кэш-модель можно рассматривать как простую n-граммную модель с вероятностями, вычисленными по предшествующей истории слов [3].

Модели, основанные на частях слов, используются в языках с богатой морфологией, например флективных языках [4], когда слово разделяется на некоторое число морфем с помощью словарных и алгоритмических методов. Преимущество алгоритмических методов в том, что они опираются лишь на анализ текста и не используют никаких дополнительных знаний, что позволяет анализировать текст на любом языке, а словарных – в том, что они позволяют получать правильное разбиение слов на морфемы, а не на псевдоморфемные единицы (как в алгоритмических методах), что может быть использовано далее на уровне пост-обработки гипотез распознавания фраз.

Модели, основанные на обобщенных синтагмах [5], под которыми понимаются контактно расположенные последовательности символов обобщенных классов словоформ, представляющих собой набор грамматических признаков слов, которые обеспечивают идентификацию в текстах слов с аналогичными грамматическими свойствами. С помощью модели обобщенных синтагм была успешно разрешена задача лексико-грамматической омонимии в английских текстах. Этому способствовал ограниченный набор грамматических признаков английских слов и их форм представления. С нашей точки зрения, дальнейшее развитие этой модели имеет большие перспективы для разрешения задач: омонимии слов в текстах; установления границ именных и глагольных словосочетаний; членения предложений на простые предложения; автоматического установления грамматически согласованной формы слова по ее контекстному окружению в решении задачи синтаксического синтеза предложений; автоматической нормализации слов и словосочетаний. Основная сложность использования этой модели для языков с развитой системой флективных окончаний (например, для русского языка) – это создание адекватного индекса обобщенной синтагмы, позволяющего отразить весь набор грамматических признаков словоформы.

СИНТАКСИЧЕСКАЯ МОДЕЛЬ ЯЗЫКА НА ОСНОВЕ ОБОБЩЕННЫХ СИНТАГМ

Концепция фразеологического концептуального анализа текстов, разработанная профессором Г.Г. Белоноговым совместно с его учениками [6], в рамках которой проводилось настоящее исследование, базируется на машинной грамматике. В основу этой грамматики положена система флективных классов русских слов [7], в которой представлены все основные типы словоизменительных парадигм лексико-грамматических классов русских слов, выявленные на основе широкомасштабных исследований лексического состава научно-технических текстов по широкому спектру тематических областей знания, а также характера изменения грамматических окончаний (флексий) русских слов и их синтаксических функций. Заложенное в теоретической концепции флективных классов слов русского языка жесткое соответствие между формой представления слов и их грамматической информацией позволило разработать на этой основе новые классы – *классы слов, имеющие одинаковые наборы грамматических признаков, соответствующие их формам представления в сходных контекстных окружениях.*

Идея создания новых классов слов, ориентированных на схожесть грамматических признаков слов и схожесть их синтаксических функций в предложении, была ранее предложена нами [5] для решения задачи разрешения грамматической омонимии английских слов. В рамках этой задачи мы исходили из следующей гипотезы: *одинаковым последовательностям обобщенных символов классов слов (обобщенным синтагмам) должны соответствовать одинаковые синтаксические структуры*. Предполагалось, что такая гипотеза верна для любых синтаксических моделей и может быть полезна при решении как глобальных, так и частных задач синтаксического анализа.

В процессе реализации предлагаемой синтаксической модели была разработана система обобщенных символов классов слов, которая отражает систему грамматических категорий слов, а также грамматические признаки конкретных форм слова, согласованных с контекстным текстовым окружением. Флективный класс слова содержит информацию о его лексико-грамматическом классе, категории рода и одушевленности (для существительных), категории возвратности (для глаголов и отглагольных форм), принадлежности к краткой или полной форме прилагательного и причастия, типе словоизменительной парадигмы, буквенном коде, длине словоизменительной основы и грамматического окончания. Сочетание двух параметров – номера флективного класса и грамматического окончания – позволяет однозначно установить все наборы основных грамматических признаков конкретной формы слова (род, число, падеж, лицо). Это сочетание было выбрано в качестве элемента предлагаемой формальной модели синтаксической структуры текстов на основе обобщенных синтагм.

По результатам анализа системы флективных классов слов и соответствующих им грамматических форм слов был сформирован двухбайтовый индекс обобщенных синтагм, состоящий из двух символов – номера флективного класса и его грамматического окончания. В табл. 1. приведены фрагменты списков буквенно-цифровых символов флективных классов и символов грамматических окончаний.

Полученные списки индексов обобщенных синтагм для каждой словоизменительной парадигмы позволили сформировать словарь соответствий между флективными классами слов, их грамматическими окончаниями, наборами грамматических признаков и индексами обобщенных синтагм (табл. 2). Этот словарь позволяет по первому символу идентифицировать грамматический класс слова, род, одушевленность (для существительных), по второму символу – грамматическое окончание и падеж.

Каждая словоизменительная парадигма в системе обобщенных синтагм может быть представлена как совокупность символа номера флективного класса и символов его грамматических окончаний. Так, символом класса 001 является символ «А» и ему соответствует совокупность следующих символов разных окончаний – {*A,B,C,D,X,G,e,n,i,q*}. Поэтому, зная индекс обобщенного класса, можно однозначно установить флективный класс слова, его грамматическое окончание и набор грамматической информации.

Таблица 1

Фрагменты списков символов флективных классов слов и грамматических окончаний

Символы флективных классов слов		Символы грамматических окончаний слов	
001 – A	*035 – d*	*+ – A*	*у – n*
002 – B	*036 – e*	*а – B*	*ут – o*
003 – C	*037 – f*	*ам – C*	*ую – p*
004 – D	*040 – g*	*ами – D*	*ы – q*
005 – E	*041 – h*	*ах – E*	*ые – r*
006 – F	*042 – i*	*ая – F*	*ый – s*
007 – G	*043 – j*	*е – G*	*ым – t*
010 – H	*044 – k*	*ев – H*	*ыми – u*
011 – I	*045 – l*	*его – I*	*ых – v*
012 – J	*046 – m*	*ее – J*	*ь – w*

Таблица 2

Фрагмент словаря соответствий между флективными классами слов, их грамматическими окончаниями, наборами грамматических признаков и индексами обобщенных синтагм

Флективный класс	Окончание	Обоб. синтагма	Грамматическая информация	Флективный класс	Окончание	Обоб. синтагма	Грамматическая информация
1	**2**	**3**	**4**	**1**	**2**	**3**	**4**
001	+	AA	1110/ 1140	103	ая	ФF	2110
001	а	AB	1120	103	ого	Фf	1120/ 3120
001	ам	AC	1230	103	ое	Фg	3110/ 3140
001	ами	AD	1250	103	ой	Фh	2120/ 2130/
001	ах	AX	1260				2150/ 2160
001	е	AG	1160	103	ом	Фi	1160/ 3160
001	ов	Ae	1220	103	ому	Фj	1130/ 3130
001	ом	Ai	1150	103	ую	Фp	2140
001	у	An	1130	103	ые	Фu	0210/ 0240
001	ы	Aq	1210/1240	103	ый	Фq	1110/ 1140
….	…	…	……..				0230/ 1150/
006	+	FA	1110/ 1140	103	ый	Фq	3150
006	а	FB	1120	103	ыми	Фu	0250
006	ам	FC	1230	103	ых	Фv	0220/ 0260
006	ами	FD	1250	….	…..	……	…….
006	ах	FX	1260	117	им	ЯV	0201
006	е	FG	1160	117	имся	ЯW	0201
006	и	FS	1210/ 1240	117	ит	ЯX	0103
006	ов	Fe	1220	117	ите	ЯY	0202
006	ом	Fi	1150	117	итесь	ЯZ	0202
006	у	Fn	1130	117	ится	Яz	0103
…..	…..	….	…….	117	ишь	Яb	0102
077	ий	HU	3220	117	ишься	Яc	0102
			3110 / 3140 /	117	ю	ЯУ	0101
077	ье	Ну	3160	117	юсь	ЯЯ	0101
077	ьем	НВ	3150	117	ят	Яу	0203
077	ью	НЛ	3130	117	ятся	Яя	0203
077	ья	НМ	3120 / 3210	……	…….	…..	…….
077	ьям	НН	/3240	141	+	уА	0000
			3230	142	+	фА	0000
077	ьями	НП	3250	143	+	цА	0000
077	ьях	НТ	3260	144	+	чА	0000
……	….		…… …	……	…..		

ПРИМЕЧАНИЕ: В первой колонке указывается флективный класс парадигмы, во второй – окончания ее словоформ, в третьей – индекс обобщенной синтагмы, в четвертой – четырехзначные наборы грамматической информации. Каждая цифра позиционного четырехзначного представления грамматической информации имеет следующее значение:

1) род: 0 – не определен, 1 – мужской, 2 – женский , 3 – средний;
2) число: 0 – не определено, 1 – единственное, 2 – множественное;
3) падеж: 0 – не определен, 1 – именительный, 2 – родительный, 3 – дательный, 4 – винительный, 5 – творительный, 6 – предложный;
4) лицо: 0 – не определено, 1 – 1-е лицо, 2 – 2-е лицо , 3 – 3-е лицо.

Пример формирования последовательности обобщенных символов синтагм предложения приведен в табл. 3, где для каждого элемента предложения на основе результатов морфологического анализа (по номеру флективного класса и буквенному коду грамматического окончания) сформирован индекс обобщенной синтагмы слова. Как видно, этот индекс в неявном виде содержит информацию о лексико-грамматическом классе слова, категории рода и одушевленности (для существительных), категории возвратности (для глаголов и отглагольных форм), принадлежности к краткой или полной форме прилагательного и причастия, типе его словоизменительной парадигмы, буквенном коде, длине словоизменительной основы и грамматического окончания. В нижней части табл. 3 показано представление предложения в виде последовательности индексов обобщенных синтагм, что, по сути, отражает *синтаксическую структуру предложения, элементами которой являются контактно расположенные объекты, обладающие грамматическими свойствами конкретных слов-эталонов. Конкретным элементам структуры могут соответствовать различные слова, грамматические признаки которых идентичны.*

В рамках теоретической концепции фразеологического концептуального анализа текстов предпола-

гается, *что устная и письменная речь имеет линейную структуру и состоит из ряда дискретных элементов – единиц смысла, в результате восприятия которой человеком в его сознании формируется некий целостный мыслительный образ. При этом в качестве единиц смысла в линейной структуре текста могут выступать единицы различного уровня: слова, словосочетания, фразы, сверхфразовые единства. Эти единицы в совокупности представляют собой иерархическую систему, в которой смысловое содержание единиц более высокого уровня не сводимо или не полностью сводимо к смысловому содержанию составляющих их единиц более низкого уровня (смысл единиц более высокого уровня не всегда может быть "вычислен" на основе информации о смысле единиц более низкого уровня и информации о связях между этими единицами). Минимальной единицей, обозначающей понятие, является слово, но большинство понятий обозначается устойчивыми словосочетаниями и фразами* [7]. Поэтому для реализации процедур автоматической обработки текстов необходимо иметь механизмы работы с этими единицами смысла, наиболее устойчивыми из которых являются слова и словосочетания, выражающие понятия. Особенно важно иметь эффективные механизмы извлечения таких словосочетаний из текстов. Сложность задачи заключается в том, что в текстах такие словосочетания формально не обозначены и только человек – специалист в данной предметной области, обладающий всей системой знаний об этой предметной области, может выделять и идентифицировать фразеологические и терминологические словосочетания-понятия. Предлагаемая авторами настоящей статьи синтаксическая модель позволит решить эту задачу в рамках следующего утверждения: *представление синтаксической структуры текстов в виде последовательности контактно расположенных двухбайтовых индексов обобщенных синтагм, обладающих грамматическими свойствами конкретных слов-эталонов, позволяет фиксировать грамматические и синтаксические свойства различных последовательностей реальных текстов, а в ряде задач – распознавать аналогичные по заданным свойствам последовательности слов и словосочетаний.*

Таблица 3

Пример представления синтаксической структуры предложения в виде последовательности символов классов обобщенных синтагм

№ n/n	Слова исходного предложения	Грамматические признаки слов	Символ обобщенной синтагмы
01	***Трубопровод-ы***	Сущ., муж. р. , неодуш., ФК=001 1)мн. ч., им. п. 2) мн. ч., вин. п.	**Aq**
02	***высок-ого***	Полн. прил., ФК=106 1)муж. р. , ед. ч, род п. 2) муж. р. , ед. ч, вин. п. 3) ср. , ед. ч, род. п.	**Цf**
03	***давлени-я***	Сущ., ср. р. ,неодуш., ФК=073 1)ед. ч., род. п. 2) мн. ч., им. п. 2) мн. ч., вин. п.	**ЙЦ**
04	***резервн-ой***	Полн. прил., ФК=103 жен. р. , ед. ч, род п. 2) жен. р. , ед. ч, дат. п. 3) жен. , ед. ч, тв. п. 4) жен. , ед. ч, пред. п.	**Фh**
05	***котельн-ой***	Полн. субстантивированное прил., ФК=103 1)жен. р. , ед. ч, род. п. 2) жен. р. , ед. ч, дат. п. 3) жен. , ед. ч, тв. п. 4) жен. , ед. ч, пред. п.	**Фh**
06	***расположен-ы***	Кратк. прич., ФК=126, мн. ч.	**жq**
07	***на***	Предлог, ФК=164, мод. упр.- вин. п., пред. п.	**7А**
08	***значительн-ом***	Полн. прил., ФК=103 1)муж. р. , ед. ч, пред. п. 2) ср. р. , ед. ч, вин. п.	**Фi**
09	***расстояни-и***	Сущ., ср. р. , неодуш., ФК=073, пред.л.	**ЙS**
10	***от***	Предлог, ФК=155, мод. упр.- род. п.	**яА**
11	***систем***	Сущ., жен. р. , неодуш., ФК=056, мн. ч., род. п.	**uА**
12	***безопасност-и***	Сущ., жен. р. ,неодуш., ФК=055 1)ед. ч., род. п. 2) ед. ч., дат. п. 3) ед. ч., пред. п. 4) мн. ч., им. п. 5) мн. ч., вин. п.	**tS**
13	***блок-а***	Сущ., муж. р. , неодуш., ФК=006, ед. ч., род. п.	**FB**
Представление предложения в виде последовательности индексов обобщенных синтагм			
Aq Цf ЙЦ Фh Фh жq 7А Фi ЙS яА uА tS FB = Трубопроводы высокого давления резервной котельной расположены на значительном расстоянии от систем безопасности блока.			

Рассмотрим несколько актуальных задач автоматической обработки текстовой информации в области информационного поиска и машинного перевода и пути их решения в рамках предлагаемой синтаксической модели текста. К таким задачам можно отнести:

- выявление в текстах именных или глагольных словосочетаний;
- автоматическая нормализация слов и словосочетаний;
- автоматическое разрешение омонимии слов;
- автоматический синтаксический анализ текстов;
- автоматическое установление структурного сходства предложений.

Выявление в текстах именных или глагольных словосочетаний

Эту задачу можно автоматически решить, используя следующую гипотезу: *если представить тексты как последовательностей обобщенных синтагм предложений и иметь информацию о синтаксической структуре именных и глагольных словосочетаний в виде последовательности обобщенных синтагм, то путем последовательного сравнения различных отрезков предложений и обобщенных синтагм эталонных словосочетаний можно выявить в текстах именные и глагольные словосочетания*. Таким образом, располагая механизмом приведения слов, словосочетаний и предложений к виду обобщенных синтагм и наличием словаря обобщенных синтагм, полученного путем приведения эталонных словосочетаний к виду обобщенных синтагм, можно достаточно эффективно решить эту задачу.

Ранее авторами был проведен эксперимент по автоматическому выявлению в текстах именных словосочетаний на основе использования словаря обобщенных синтагм эталонных словосочетаний. В качестве исходных данных для этого эксперимента был выбран корпус общественно-политических текстов общим объемом 10 Мб, предварительно обработанный комплексом процедур автоматизированного составления словарей наименований понятий по текстам документов [8]. В результате был получен частотный словарь словосочетаний. Фрагмент частотного словаря именных словосочетаний, в котором обобщение текстовых форм представления словосочетаний производилось на уровне их словоизменительных парадигм приведен в табл. 4. Для каждого словарного словосочетания указана его частота в корпусе текстов, нормальная форма на уровне словоизменения и одна из текстовых форм представления этого словосочетания в тексте.

Для формирования словаря обобщенных синтагм была использована частотная часть словаря именных словосочетаний. Автоматизированное составление словаря обобщенных синтагм на основе использования словаря именных словосочетаний можно выполнить по следующей технологической схеме.

Этап 1. Обработка именных словосочетаний словаря по процедуре морфологического анализа.

Этап 2. Построение для каждого словосочетания его обобщенной синтагмы.

Этап 3. Формирование частотного словаря обобщенных синтагм.

Этап 4. Лингвистический анализ частотного словаря обобщенных синтагм (исключение ошибочных и малоинформативных синтагм).

Этап 5. Формирование машинного представления словаря обобщенных синтагм.

Результаты обработки словосочетаний по этой схеме были представлены в виде частотного словаря обобщенных синтагм, соответствующих синтаксическим структурам словосочетаний-эталонов. Фрагмент частотного словаря обобщенных синтагм именных словосочетаний приведен в табл. 5. Следует отметить, что в этом словаре обобщение выполнялось на уровне синтаксических структур словосочетаний-эталонов. В частотном словаре для каждой словарной статьи указывалась частота встречаемости синтаксической структуры словосочетания-эталона в корпусе текстов, обобщенная синтагма в виде последовательности индексов обобщенных классов слов и эталонное словосочетание – одна из текстовых форм словосочетаний, соответствующая обобщенной синтагме.

Таблица 4

Фрагмент частотного словаря именных словосочетаний

00000134 соединенный штат * Соединенные Штаты	00000023 природный газ * природного газа
00000079 советский союз * Советский Союз	00000022 второй мировой война * Второй мировой войне
00000053 холодный война * Холодная война	00000022 цена на нефть * цен на нефть
00000052 ближний восток * Ближнего Востока	00000020 саудовский аравия * Саудовская Аравия
00000046 иностранный дело * иностранных дел	00000019 боевой действие * боевые действия
00000040 вооруженный сила * Вооруженные Силы	00000019 противоракетный оборона * противоракетной обороне
00000038 персидский залив * Персидский залив	00000018 международный отношение * Международным отношениям
00000036 ядерный оружие * Ядерное оружие	00000018 официальный лицо * официальные лица
00000035 северный корея * Северная Корея	00000017 распад советский союз * Распад Советского Союза
00000033 ядерный программа * ядерная программа	00000017 сельский хозяйство * сельского хозяйства
00000032 мировой война * мировая война	00000016 арабский мир * арабский мир
00000031 олимпийский игра * Олимпийские игры	00000016 баллистический ракета * баллистические ракеты
00000028 гражданский война * Гражданская война	00000016 белый дом * Белого Дома
00000028 совет безопасность * Совет Безопасности	
00000025 национальный безопасность * национальной безопасности	
00000025 поставка оружие * Поставка оружия	
00000021 подводный лодка * Подводные лодки	

Таблица 5

Фрагмент частотного словаря обобщенных синтагм именных словосочетаний

00000429 ФsAA / Апелляционный суд	00000137 ФhtS / Восточной Сибири
00000312 ФgЙG / 35-страничное заявление	00000124 ФFuB / Бубонная чума
00000284 ФfAB / 11-месячного конфликта	00000121 ФhuG / Восточной Европе
00000277 ФrAq / Авторитарные режимы	00000114 ЦhxS / Американской демократии
00000257 Фf ЙЦ / Бюджетного управления	00000111 ФtAi / Государственным Советом
00000224 ЦUAA / Азиатско-Тихоокеанский регион	00000111 ФvЙс / авторитарных устремлений
00000184 ФhxS / Белой гвардии	00000108 Фpun / агрессивную войну
00000176 ФvAe / Положительных отзывов	00000106 ФgЖd / Всемирное братство
00000164 ЦfAB / Американского совета	00000099 ЦgЙG / Берлинское отделение
00000162 Фruq / Вооруженные Силы	00000096 ЦfЙЦ / Верхнечонского месторождения

Таблица 6

Результаты работы процедуры автоматического выявления именных словосочетаний в корпусе текстов

1	**Обобщенная синтагма именного словосочетания = ФsAA** **Частота встречаемости в корпусе текстов = 429** **Текстовые словосочетания, соответствующие обобщенной синтагме;** Анонимный блогер, 30-градусный мороз, 569-страничный документ, Апелляционный суд, Атомный ледокол, Безмешковый пылесос, Винный мир, Властный вакуум, Военный бюджет, Военный вариант, Гендерный дисбаланс, Глобальный фонд, Горный массив, Западный вектор, Западный проект, Зеленый свет…
2	**Обобщенная синтагма именного словосочетания = ФgЙG** **Частота встречаемости в корпусе текстов = 348** **Текстовые словосочетания, соответствующие обобщенной синтагме;** Изначальное пожертвование, Классовое сознание, Кумулятивное воздействие, Машинное обучение, Новое мышление, Персонифицированное обучение, Полное молчание, Программное обеспечение, Финальное движение, Эффективное управление, Ядерное оружие, авторитарное правление, агрессивное выступление, адаптивное поведение, акционерное соглашение, алкогольное отравление, атомное оружие, безобидное голосование, бесперебойное обеспечение, беспорядочное вращение…
3	**Обобщенная синтагма именного словосочетания = ФhuG** **Частота встречаемости в корпусе текстов = 256** **Текстовые словосочетания, соответствующие обобщенной синтагме;** Восточной Европе, Государственной Думе, Западной Европе, Красной Поляне, Красной планете, Независимой газете, Силиконовой долине, атомной сфере, бесконечной зиме, взаимной обороне, военной базе, военной борьбе, военной диктатуре, военной карьере, военной пропаганде, военной тюрьме, военной форме, воздушной обороне, вольной борьбе, всемирной паутине, газопроводной системе, глобальной войне, государственной измене, двухкомнатной квартире, действительной службе.

Автоматическое выявление именных словосочетаний в корпусе текстов мы выполняли в соответствии с алгоритмом 1.

Алгоритм 1. Автоматическое выявление именных словосочетаний в тексте.

***Шаг* 1.** Разделить текст на предложения, с указанием позиций начала и длины каждого предложения;

***Шаг* 2.** Выделить в каждом предложении все возможные его фрагменты с указанием позиций начала и длины каждого фрагмента предложения;

***Шаг* 3.** Каждый фрагмент предложения обработать процедурой морфологического анализа и по результатам этого анализа построить обобщенную синтагму фрагмента;

***Шаг* 4.** Каждую сформированную синтагму фрагмента сопоставить с синтагмами словаря обобщенных синтагм, и в случае их совпадения считать этот фрагмент предложения именным словосочетанием;

***Шаг* 5.** Сформировать машинное представление полученных метаданных с указанием в них выделенных словосочетаний (в исходной и нормальной форме), их позиции в тексте (в символах) и длины.

Результаты работы процедуры автоматического выявления именных словосочетаний в корпусе текстов приведены в табл. 6, где для большей информативности указываются обобщенная синтагма именного словосочетания, частота встречаемости в корпусе текстов, фрагмент списка словосочетаний, соответствующих структуре обобщенной синтагмы.

Автоматическая нормализация слов и словосочетаний

Эту задачу необходимо разбить на две подзадачи: нормализации слов и нормализации словосочетаний. Нормализация слов русского языка на различных уровнях обобщения достаточно хорошо исследована в работах [6, 7] и решается путем замены грамматических окончаний, а в некоторых случаях – трансформации конечных буквосочетаний основ. Другое дело – нормализация именных и глагольных слово-

сочетаний. Здесь, наряду с задачей морфологического синтеза, необходимо проанализировать структуру текстового словосочетания: установить главные (опорные) и зависимые слова и определить синтаксические связи между ними. На основе этой информации может быть принято решение о синтаксической и морфологической трансформации опорных и зависимых слов. Под синтаксической трансформацией понимается изменение порядка следования слов в нормализующем словосочетании или изменение грамматических классов слов (например, трансформация формы существительного в форму прилагательного). Под морфологической трансформацией понимается изменение грамматического окончания слова или, при необходимости, трансформация буквенного состава основы слова. Эти процедуры разработаны, но их функционирование требует значительных вычислительных и временных ресурсов. Механизм обобщенных синтагм позволит решить эту задачу гораздо эффективнее с помощью метода лингвистической аналогии, базирующегося на гипотезе: *одним и тем же синтаксическим структурам словосочетаний должны соответствовать одинаковые синтаксические структуры их нормальных форм. При этом все процедуры преобразования в нормальные формы выполняются по упрощенному варианту формирования нормальных представлений словосочетаний.*

Таким образом, задачу автоматической нормализации словосочетаний можно свести к нахождению синтаксической структуры в виде обобщенной синтагмы словаря, аналогичной анализируемому словосочетанию, и простого механизма трансформации в нормальную форму путем замены грамматических окончаний слов словосочетаний. Но такая трансформация относится только к простым случаям нормализации. Что касается сложных случаев нормализующих трансформаций, то здесь уже необходимо либо выполнить перестановку слов, либо, наряду с изменением окончания слов, потребуется осуществить соответствующую трансформацию их основ. Но таких сложных случаев трансформаций словосочетаний относительно немного, поэтому их можно включить в словарь нормализующих трансформаций словосочетаний (НТС). При этом составление такого словаря можно выполнить автоматически с лингвистическим контролем. В настоящее время этот словарь имеет объем 39865 словарных статей. В табл. 7 приведены результаты работы процедуры автоматической нормализации слов и словосочетаний.

В ряде задач автоматической обработки текстовой информации требуется большая степень обобщения словосочетаний с одним и тем же смысловым содержанием, а также возможность поиска словосочетаний по их частичному нормализованному представлению, например, по главным словам словосочетаний. В этом случае нормализация выполняется путем пословной нормализации и изменения порядка следования слов в нормализованном словосочетании. Поэтому в рамках этой версии указывался порядок следования слов в нормализованном словосочетании. В табл. 8 приведены результаты работы процедуры автоматической унификации слов и словосочетаний.

В соответствии с описанной методикой автоматической нормализации словосочетаний был разработан алгоритм 2.

Алгоритм 2. Автоматическая нормализация словосочетаний.

***Шаг* 1.** Выполнить обработку анализируемого словосочетания с помощью морфологического анализа и построить обобщенную синтагму словосочетания.

***Шаг* 2.** Произвести поиск полученной обобщенной синтагмы в словаре НТС. В случае успешного поиска такой синтагмы исходное словосочетание заменить по словарю на его нормальную форму и выполнить переход к шагу 5.

***Шаг* 3.** Произвести поиск обобщенной синтагмы в словаре НТС. В случае успешного поиска выполнить нормализацию слов словосочетания путем замены исходных грамматических окончаний каждого слова словосочетания на их нормализующие окончания в соответствии с левым символом каждого индекса слова словосочетания. При удачном поиске – переход к шагу 5.

***Шаг* 4.** Выполнить словоизменительную нормализацию словосочетания традиционным способом [6].

***Шаг* 5.** Преобразовать полученные результаты в структуру метаданных.

Таблица 7

Результаты работы процедуры автоматической нормализации слов и словосочетаний

Исходная синтагма	Нормализующая синтагма	Исходное словосочетание	Нормализованное словосочетание
ФuAD	ФrAq	*Соединенными Штатами*	*Соединенные Штаты*
ЦVAi	ЦUAA	*Советским Союзом*	*Советский Союз*
ФvuE	Фruq	*Вооруженных Силах*	*Вооруженные Силы*
ФiЙL	ФgЙG	*Ядерном оружием*	*ядерное оружие*
ФhyG	ФFyЦ	*Северная Корея*	*Северная Корея*
Фhuh	ФFuB	*ядерной программой*	*ядерная программа*
ФhtS	ФhtS	*национальной безопасности*	*национальная безопасность*
ЧhЧhuG	ЧhЧhuG	*Второй мировой войне*	*Вторая мировая война*
ФhuG1AFn-6АЙLxc	ФFuB1AFn-6АЙLxc	*Международной службе по мониторингу за применением агробиотехнологий*	*международная служба по мониторингу за применением агробиотехнологий*
ФрЦpun	ФFЦFuB	*межконтинентальную баллистическую ракету*	*межконтинентальная баллистическая ракета*

Таблица 8

Результаты работы процедуры автоматической унификации слов и словосочетаний

Исходная синтагма	Нормализующая синтагма	Нормализующие перестановки	Исходное словосочетание	Нормализованное словосочетание
ФrAq	AAФs	*0201*	*Соединенные Штаты*	*штат соединенный*
ЦUAA	AAЦU	*0201*	*Советский Союз*	*союз советский*
Фruq	uBФs	*0201*	*Вооруженные Силы*	*сила вооруженный*
ФgЙG	ЙGФs	*0201*	*Ядерное оружие*	*оружие ядерный*
ФFyЦ	yЦФs	*0201*	*Северная Корея*	*корея северный*
ФFuB	uBФs	*0201*	*ядерная программа*	*программа ядерный*
ФhtS	twФs	*0201*	*национальной безопасности*	*безопасность национальный*
ЧhЧhuG	uBЧhЧh	*030201*	*Второй мировой войне*	*война второй мировой*
ФFuB1AFn-6АЙLxc	uBФsFАЙGxЦ	*0201040506*	*Международная служба по мониторингу за применением агробиотехнологий межконтинентальную баллистическую ракету*	*служба международный мониторинг применение агробиотехнология*
ФрЦpun	uBЦUФs	*030201*		*ракета баллистический межконтинентальный*

Автоматическое разрешение омонимии слов

Известно, что в естественных языках распространено такое явление как омонимия слов. Автоматическое разрешение грамматической омонимии для английского языка описано в работе [5], где на основе исследований лексического состава и синтаксической структуры английских текстов было установлено, что грамматическая омонимия английских многозначных слов может быть разрешена только с помощью контекста, а контекст в обобщенном виде может быть представлен в виде последовательностей символов грамматических классов слов – обобщенных синтагм, а также что для каждого слова с неоднозначной грамматической информацией должен быть приведен микроконтекст из нескольких элементов текста и указана однозначная информация, соответствующая этому контексту. Общим принципом установления однозначной грамматической информации с помощью обобщенных синтагм может быть гипотеза: *одинаковым обобщенным синтагмам соответствует одинаковая однозначная грамматическая информация к словам, стоящим в центре отрезков текста, которые эти синтагмы представляют.*

Этот принцип был принят за основу при разработке метода разрешения лексико-грамматической омонимии русских слов, который базируется на результатах морфологического анализа слов и процедуре установления однозначной информации по контексту. В табл. 9 приведены результаты морфологического анализа омонимичного слова «*суда*».

Для разрешения лексико-грамматической многозначности слова необходимо использовать его контекстное окружение, представленное в виде последовательности контактно расположенных слов предложения, предшествующих данному слову и последующих за ним. Отсутствие слов в этой последовательности обозначается двухсимвольным идентификатором «*ZZ*». Для поиска в словаре контекстных синтагм (КС) полученная последовательность синтагм преобразовывалась в модифицированную (поисковую) контекстную синтагму. В табл. 10 показано преобразование контекстного окружения омонимичного слова «суда» в его контекстную поисковую синтагму.

Далее производится поиск в словаре СК на наибольшее вхождение символов анализируемой синтагмы и символов синтагмы словаря. Однозначная грамматическая информации устанавливается в соответствии с грамматической информацией этой синтагмы. На базе описанной методики автоматического разрешения лексико-грамматической омонимии слов был разработан алгоритм 3.

Алгоритм 3. Автоматическое разрешение лексико-грамматической омонимии слов

***Шаг* 1.** Разделить текст на предложения с указанием позиций начала и длины каждого предложения.

***Шаг* 2.** Каждое предложение обработать процедурой морфологического анализа и по результатам этого анализа установить случаи, требующие разрешения лексико-грамматической омонимии слов.

***Шаг* 3.** Построить для каждого омонимичного слова предложения поисковую синтагму.

***Шаг* 4.** Произвести поиск полученной поисковой синтагмы в словаре СК на наибольшее совпадение символов поисковой и словарной синтагм.

***Шаг* 5.** Проверить наличие полученной по словарю грамматической информации и информации, полученной в результате морфологического анализа. В случае успешного обнаружения такой информации назначить слову эту однозначную лексическую информацию. Если такая информация отсутствует, то назначить наиболее близкую по грамматическому классу и типу словоизменения информацию из имеющейся в наборе грамматической информации класса слова.

***Шаг* 6.** Выполнить преобразование полученных результатов в структуру метаданных.

Таблица 9

Результат морфологического анализа омонимичного слова

Текстовое слово	Флективный класс №1	Флективный класс №2	Окончание	Индекс синтагмы №1	Индекс синтагмы №2	Грамматическая информация №1	Грамматическая информация №2
суда	*001*	*070*	*а*	*AB*	*fB*	*3210/3240*	*1120*

Таблица 10

Пример формирования модифицированной (поисковой) синтагмы омонимичного слова по контексту предложения

Перенумерованное исходное предложение №1	
01#Большегрузные 02#суда 03#стояли 04#на 05#открытом 06#рейде	
Процесс трансформации синтагмы	
01 02 03 04 05 06	Нумерация исходной синтагмы
Фr AB дS 7A Фi AG	Исходная синтагма
-- Ом -- -- -- --	Признак наличия омонимии
-5 -4 -3 -2 -1 00 +1 +2 +3 +4 +5	Нумерация контекста синтагмы
ZZ ZZ ZZ ZZ Фr XX дS 7A Фi AG ZZ	Контекст синтагмы
00 01 02 03 04 05 06 07 08 09 10	Нумерация модифицированной синтагмы
XX Фr дS ZZ 7A ZZ Фi ZZ AG ZZ ZZ	Модифицированная синтагма
ФrдSZZ7AZZФiZZAGZZZZ	Поисковая синтагма
ФrдSZZ7AZZФiZZAGZZZZ=AB	Найденная синтагма в словаре КС

Перенумерованное исходное предложение №2	
01#Заседание 02#Таганского 03# суда 04#было 05#перенесено 06#на 07#следующую 08#неделю	
Процесс трансформации синтагмы	
01 02 03 04 05 06 07 08	Нумерация исходной синтагмы
ЙG Цf AB дd жd 7A 8р zУ	Исходная синтагма
-- -- Ом -- -- -- -- --	Признак наличия омонимии
-5 -4 -3 -2 -1 00 +1 +2 +3 +4 +5	Нумерация контекста синтагмы
ZZ ZZ ZZ ЙG Цf XX дd жd 7A 8р zУ	Контекст синтагмы
00 01 02 03 04 05 06 07 08 09 10	Нумерация модифицированной синтагмы
XX Цf дd ЙG жd ZZ 7A ZZ 8р ZZ zУ	Модифицированная синтагма
ЦfдdЙGжdZZ7AZZ8рZZzУ	Поисковая синтагма
ЦfдdЙGжdZZ7AZZ8рZZzУ=fB	Найденная синтагма в словаре КС

Автоматический синтаксический анализ текстов

Основная задача синтаксического анализа – представить синтаксическую структуру предложений в терминах классов слов и их отношений. При этом в качестве классов слов могут выступать части речи (существительное, прилагательное, глагол, наречие и др.), сопровождаемые грамматической информацией, характеризующей конкретные формы слов (например, род, число, падеж, лицо и др.); в качестве отношений – отношения непосредственной доминации с той или иной степенью их дифференциации.

Исходными данными для реализации процедуры традиционного синтаксического анализа, основанного на правилах, являются результаты морфологического анализа слов и набор правил конкретного естественного языка. Для русского языка с помощью этих правил и грамматической информации, полученной на этапе морфологического анализа, строятся различные синтаксические модели текста, а также устанавливается его предикатно-актантная структура, которая, по сути, является смысловым инвариантом, обеспечивающим возможность автоматического перевода текстов с естественных языков на формализованные и с формализованных на естественные [7].

В процессе синтаксического анализа для каждого предложения последовательно решаются его конкретные задачи: а) определяются грамматические классы слов предложения; б) устанавливаются границы простых предложений; в) устанавливаются границы глагольных и именных словосочетаний и определяются в них главные слова; г) формируется «скелет» предложения (последовательность главных слов словосочетания, предлогов, союзов и знаков препинания); д) устанавливается однозначная грамматическая информация слов по их контексту; е) определяются синтаксические связи между словами предложения; ж) устанавливаются главные и второстепенные члены предложения; з) строится «дерево зависимостей» предложения; и) выявляется предикатно-актантная структура (ПАС) простого предложения («субъект» (*S*), «объект» (*O*) и «предикат» (*P*)). В табл. 11 представлен результат обработки предложения этими операциями процедуры синтаксического анализа.

Таблица 11

Результат работы процедуры синтаксического анализа текстов

Перенумерованное исходное предложение		
01#Трубопроводы 02#высокого 03#давления 04#резервной 05#котельной 06#расположены 07#на 08#значительном 09#расстоянии 10#от 11#систем 12#безопасности 13#блока.		
Результаты синтаксического анализа предложения		
Мнемоническое обозначение	**Формальная структура предложения**	**Пояснения к элементам формальной структуры предложения**
Num1	*0000000000111*	Нумерация слов в предложении (символы нумерации расположены вертикально)
Num2	*1234567890123*	
Frm0	*NANAAKFANFNNN.*	Предложение в виде символов грамматических классов слов
Snt1	*АЦЙФФж7ФЙяutF*	Предложение в виде символов обобщенных синтагм (синтагмы расположены вертикально)
Snt2	*qfUhhqAiSAASB*	
Bnd0	*( )||()|( )*	Границы слов и словосочетаний в предложении
Skl0	*S PF NFN*	«Скелет» предложения в символах классов слов
SPO	*SSSSSPPOO OOO*	Позиции элементов ПАС предложения
Frm0	*NANAAKFANFNNN.*	Предложение в виде символов грамматических классов слов
Lnk1	*0000000000111*	Дерево зависимостей предложения (для каждого слова «слуги» указывается номер позиции слова «хозяина»)
Lnk1	*6315116966612*	
Gnd	*1332210330221*	Род слова (установленный по контексту)
Nmb	*2111120110211*	Число слова (установленное по контексту)
Case	*1222202222222*	Падеж слова (установленный по контексту)
Fase	*0000000000000*	Лицо слова (установленное по контексту)
SPO1	*NANAA-KF-AN*	ПАС предложения №1 в символах классов слов
	AqЦfЙЩФhФh-жqяA-ФiЙS	ПАС предложения №1 в символах обобщенных синтагм
	01,05-06,02-09-02	Позиции первых слов конструкций ПАС предложения №1 и их длин
SPO2	*NANAA-KF-NNN*	ПАС предложения №2 в символах классов слов
	AqЦfЙЩФhФh-жqяA-uAtSFB	ПАС предложения №2 в символах обобщенных синтагм
	01,05-01:10,01-11,03	Позиции первых слов конструкций ПАС предложения №2 и их длин

№ ПАС	***Текстовое представление ПАС***	***Формализованное представление ПАС***
1	*Трубопроводы высокого давления резервной котельной - расположены на - значительном расстоянии*	*Трубопровод - располагать на- расстояние*
2	*Трубопроводы высокого давления резервной котельной - расположены от - систем безопасности блока*	*Трубопровод - располагать от - система*

Очевидно, что традиционный алгоритм синтаксического анализа не может обеспечить достаточного быстродействия при обработке больших массивов текстов. При этом также известно, что опираясь только на общие синтаксические правила, не всегда возможно правильно построить синтаксические структуры текстов при автоматическом анализе алгоритмически неразрешимых текстовых ситуаций. Поэтому в некоторых случаях функционирование таких алгоритмов не всегда достигает желаемого результата. Обычно повышение качества обработки текстов решается с помощью привлечения дополнительных семантических признаков слов и использования шаблонов, моделирующих структуру синтаксических конструкций предложений.

Одним из наиболее эффективных методов решения этой проблемы может быть разработка алгоритмов синтаксического анализа текстов, ориентированных на машинное обучение. Такое обучение можно также реализовать на основе предлагаемой авторами модели, учитывая тот факт, что процедуры синтаксического анализа базируются на результатах морфологического анализа текста, которые возможно автоматически преобразовать в последовательность индексов обобщенных синтагм, однозначно соответствующих конкретной синтаксической структуре предложения (см. табл. 4). При этом, если такой последовательности синтагм предложений-эталонов поставить в соответствие ранее выполненные результаты их обработки традиционным алгоритмом синтаксического анализа, а также обеспечить возможность ручного исправления в них алгоритмически неразрешимых некорректностей, то можно путем соотнесения текстовых синтагм и синтагм словаря, в котором заранее получена

и откорректирована синтаксическая информация, получить качественное решение алгоритмически неразрешимых задач и на этой основе разработать эффективный и высококачественный автомати-ческий синтаксический анализ текстов.

Решение этой задачи осложняется большой вариативностью текстового представления синтаксических структур предложений, особенно учитывая тот факт, что в русском языке порядок слов в них относительно произволен и одна и та же синтаксическая структура может быть выражена различными текстовыми представлениями. Поэтому основная проблема создания словарного ресурса, содержащего информацию об синтаксических структурах предложений, будет заключаться в обеспечении возможности определения тождественных или близких по семантико-синтаксической структуре различных текстовых представлений предложений.

Автоматическое установление структурного сходства предложений

Такое определение тождественных или близких по семантико-синтаксической структуре различных текстовых представлений предложений должно базироваться на процедуре автоматического установления сходства синтаксических структур различных предложений, позволяющей использовать результаты анализа заранее обработанных и верифицированных предложений-эталонов. Принимая во внимание, что в русском языке порядок следования синтаксических конструкций в предложении однозначно не определен, а смысловая связь между элементами предложения устанавливается позициями слов в этих конструкциях и грамматическими окончаниями слов предложения, то при создании такой процедуры нужно опираться на формальное представление синтаксической структуры предложения, полученное на этапе его традиционного синтаксического анализа. При этом необходимо использовать результаты формализации синтаксической структуры предложения и его структурных элементов, которые позволят повысить распознающую способность этой процедуры. Такую формализацию можно представить в виде набора шаблонов, отражающих ее основные структурные конструкции: «скелет» предложения, шаблоны именных и глагольных словосочетаний, шаблоны причастных и деепричастных оборотов и другие шаблоны синтаксических конструкций предложения. Для решения этой задачи необходимо на основе предлагаемых решений создать словарь структур эталонных предложений (СЭП) на базе обобщенных синтагм и дополнительно к нему несколько словарей эталонных синтаксических конструкций, в частности, словарь структур эталонных словосочетаний (СЭС), структур причастных и деепричастных оборотов и т.д.

Для поиска в этом словаре по таким же принципам следует создать формализованное представление структуры анализируемого предложения, используя при этом упрощенную процедуру синтаксического анализа, основной функцией которого будет определение конструкций подлежащего и сказуемого и их главных слов, а также конструкций второстепенных членов предложения – прямого или косвенного дополнения, обстоятельства места, времени и т.д. На основе этой информации можно построить несколько модифицированных представлений анализируемого предложения – модифицированный (поисковый) «скелет» предложения на базе символов классов и символов обобщенных синтагм слов с указанием позиций членов предложения и формализованное представление синтаксических конструкций, входящих в состав анализируемого предложения. Пример такой формализации был приведен в табл. 11.

Результаты упрощенного преобразования текстового представления предложения в модифицированное (поисковое) представление его синтаксической структуры, построенной на основе использования обобщенных синтагм показаны в табл. 12. В верхней части этой таблицы приведено текстовое представление с нумерацией каждого слова предложения, в средней части – показаны результаты поэтапного преобразования текстовой формы предложения в его формальное представление, и в нижней части – показана формальная семантико-синтаксическая структура предложения на основе обобщенных синтагм. В этой структуре вся информация позиционно разделена на блоки грамматической, синтаксической и семантической информации. Блоки информации в словарной статье представлены в позиционной форме и разделяются идентификаторами следующего вида: номер блока окаймлен с обеих сторон знаком «=» (равно), информация внутри блоков разделяется знаком «–» (тире) или знаком «#» (решетка), однородная информации следует через запятую. Блоки информации расположены в следующем порядке: 01 – поисковая синтагма предложения (модифицированный «скелет» предложения); 02 – формализованное текстовое представление «скелета» предложения; 03 – поисковая синтагма предложения с указанием позиции главного слова словосочетания в предложении и его идентификатора; 04 – дерево зависимостей «скелета», представленного в символах обобщенных синтагм; 05 – модифицированный «скелет» предложения в символах классов слов с указанием дерева зависимостей (в терминах «хозяин-слуга») в словосочетании, набора однозначной грамматической информации (род, число, падеж, лицо) для каждого слова словосочетания и количества слов «скелета»; 06 – число словосочетаний в предложении; 07 – информация о каждом словосочетании предложения: структура словосочетаний в символах классов слов и символах обобщенных синтагм с указанием позиции первого слова в предложении (для разрывных конструкций через запятую указан номер позиции второй части словосочетания), позиции главного слова в предложении и в словосочетании и его длины; 08 – число ПАС; 09 – информация о конструкции ПАС в виде символов обобщенных синтагм словосочетаний и их позиций в предложении (позиция первого слова словосочетания в предложении и его длина); 10 – текстовое представление ПАС; 11 – структура эталонного предложения в символах обобщенных синтагм, символах классов слов и длины предложения; 12 – эталонное предложение в текстовом представлении.

Таблица 12

Результаты преобразования текстового представления предложения в его формальную модель на основе обобщенных синтагм

Перенумерованное исходное предложение		
01#Трубопроводы 02#высокого 03#давления 04#резервной 05#котельной 06#расположены 07#на 08#значительном 09#расстоянии 10#от 11#систем 12#безопасности 13#блока.		
Результаты упрощенного синтаксического анализа предложения		
Мнемонические .обозначения	**Формальная структура предложения**	**Пояснения к элементам Формальной структуре предложения**
Num1	*0000000000111*	Нумерация слов в предложении (символы нумерации расположены вертикально).
Num2	*1234567890123*	
Frm0	*NANAAKFANFNNN*	Предложение в виде символов грамматических классов слов.
Snt1	*АЦЙФФж7ФЙяutF*	Предложение в виде символов обобщенных синтагм (синтагмы расположены вертикально).
Snt2	*qfЦhhqAiSAASB*	
Skl0	*S PF NFN*	«Скелет» предложения в символах классов слов.
Построение формализованного представления «скелета» предложения		
Num0	*123456*	Нумерация слов в модифицированном «скелете» предложения.
Skl1	*KNFNFN*	Модифицированный (поисковый) «скелет» предложения на основе символов классов слов.
Lnk1	*123456*	Дерево зависимостей предложения (для каждого слова «слуги» указывается номер позиции слова «хозяина»).
Lnk1	*211315*	
Gnd	*110302*	Информация о роде слов «скелета».
Nmb	*220102*	Информация о числе слов «скелета».
Case	*012222*	Информация о пдеже слов «скелета».
Fase	*000000*	Информация о лице слов «скелета».
WC01	*NANAN-AqЦfЙЦФhФh-01-05*	Структура словосочетаний в символах классов слов и символах обобщенных синтагм с указанием позиции первого слова в предложении (для разрывных конструкций через запятую указан номер позиции второй части словосочетания) и длины словосочетания. Нумерация словосочетаний указана в порядке их следования в предложении.
WC02	*NAN-AqЦfЙЦ-01-03*	
WC03	*AA-ФhФh-04-02*	
WC04	*KF-жq7A-06-02*	
WC05	*KF-жqяA-06,10-02*	
WC06	*AN-ФiЙS-08-02*	
WC07	*NNN-uAtSFB-11-03*	
WC08	*NN-uAtSFB-12-02*	
Skl2	*KNFNFN-«располагать-трубопровод–на-расстояние - от-система»*	Модифицированный «скелет» предложения в символах классов слов и его формализованное семантическое представление.
Skl3	*ЖqAqAЙSяAuuA*	Модифицированный (поисковый) «скелет» предложения в символах обобщенных синтагм
Skl4	*Жq-06-WC03#Aq-01-WC01#ЙS-09-WC05#uA-11-WC06*	Модифицированный «скелет» предложения в символах обобщенных синтагм с указанием позиции главного слова словосочетания в предложении и их идентификаторами.
Формальная структура предложения на основе обобщенных синтагм		
***01=ЖqAqAЙSяAuuA=02**=«располагать-трубопровод–на-расстояние-от-система»=**03**=Жq-06-WC04#Aq-01-WC01#ЙS-09-WC065#uA-11-WC06=**04**=020101030105=**05**=KNFNFN-PSFDFD-110302-220102-012222-000000-06=**06**=NWC-8=**07**=WC01-NANAN-AqЦfЙЦФhФh-01-01-05=WC02-NAN-AqЦfЙЦ-01-01-03=WC03-AA-ФhФh-04-02-02#WC04-KF-жq7A-06-01-02#WC05-KF-жqяA-06,10-01-02#WC06-AN-ФiЙS-08-02-02#WC07-NNN-uAtSFB-11-01-03#WC08-NN-tSFB-12-01-02=**08**=NSPO-2=**09**=AqЦfЙЦФhФh-жq7A-uAtSFB#01,05-06,02-08,02##AqЦfЙЦФhФh-жqяA-uAtSFB#01,05-06,02-08,02-11=**10**=«трубопровод–располагать-на-расстояние»=«трубопровод–располагать-от-система»=**11**=AqЦfЙЦФhФhжq7AФiЙSяAuAtSFB-NANAAKFANFNNN-13=**12**=«Трубопроводы высокого давления резервной котельной расположены на значительном расстоянии от систем безопасности блока»*		

Далее необходимо построить формализованное представление каждой синтаксической конструкции предложения. Проиллюстрируем это на примере формализации именных и глагольных словосочетаний анализируемого предложения, приведенном в табл. 13. В верхней части этой таблицы приведены все словосочетания предложения с указанием для каждого словосочетания его идентификатора, исходной (текстовой) и нормализованной формы представления, позиции словосочетания в предложении, его

типа (именное – *N* или глагольное – *V*) и его длины; в средней части – приведены результаты синтаксического анализа словосочетания. Такими результатами для каждого словосочетания являются: структура словосочетания в символах обобщенных синтагм и в символах классов слов с указанием дерева зависимостей (в терминах «хозяин слуга») в словосочетании; набор однозначной грамматической информации (род, число, падеж, лицо) для каждого слова словосочетания; тип внутренней и внешней синтаксической связи в словосочетании в символах обобщенной синтагмы и символах классов слов. И наконец, в нижней части этой таблицы информация о синтаксической и семантической структуре словосочетания представлена в позиционной форме в следующем порядке: 1) структура словосочетания в символах обобщенных синтагм; 2) символ главного слова; 3) символ внешнего управляющего слова; 4) позиция главного слова в словосочетании; 5) тип словосочетания; 6) дерево зависимостей (главное слово представлено символом обобщенной синтагмы, для зависимых слов указывается номер позиции «хозяина»); 7) структура словосочетания в символах обобщенных синтагм; 8) информация о роде слов словосочетания; 9) информация о числе слов словосочетания; 10) информация о падеже слов словосочетания; 11) информация о лице слов словосочетания; 12) формализованное и текстовое представление словосочетания.

Словарная статья словаря СЭП идентична формальному описанию структуры предложения, представленной в нижней части табл. 12, с тем отличием, что внутренние идентификаторы словосочетаний предложений будут заменены на их уникальные номера словаря СЭС, как видно из примера, приведенного в табл. 14.

В соответствии с предложенной методикой были разработаны алгоритмы формирования словаря СЭС и словаря СЭП, для которых в качестве исходных данных должны быть использованы репрезентативные корпуса отраслевых научно-технических текстов.

Алгоритм 4. Автоматическое формирование словаря СЭС

***Шаг* 1. Р**азделить текст на предложения и выполнить обработку каждого анализируемого предложения процедурой морфологического анализа.

***Шаг* 2.** Произвести синтаксический анализ предложения: определить границы словосочетаний, главные и второстепенные члены предложения, построить дерево зависимостей предложения и сформировать «скелет» предложения.

***Шаг* 3.** Выделить в анализируемом предложении именные и глагольные словосочетания.

***Шаг* 4.** Каждому словосочетанию присвоить его уникальный идентификатор в предложении.

***Шаг* 5.** Для каждого словосочетания определить его позицию в предложении и длину, построить его обобщенную синтагму, установить главное слово и индекс его синтагмы.

***Шаг* 6.** Для главного слова словосочетания определить его позицию в словосочетании и установить его «внешнее» управление (его «хозяин» в предложении) и индекс синтагмы этого управляющего слова.

Шаг 7. Внутри словосочетания определить все зависимые слова и их связи типа «хозяин-слуга», построить дерево зависимостей (главное слово представлено символом обобщенной синтагмы, для зависимых слов указать номер позиции «хозяина»).

***Шаг* 8.** Для каждого слова словосочетания установить его однозначную грамматическую информацию (род, число, падеж, лицо).

***Шаг* 9.** Произвести автоматическую нормализацию исходного словосочетания.

***Шаг* 10.** По нормализованному представлению словосочетания построить его нормализованную синтагму.

***Шаг* 11.** Каждому элементу «скелета», являющемуся главным словом словосочетания, поставить в соответствие идентификатор его словосочетания.

***Шаг* 12.** По результатам этой обработки построить словарную статью словаря СЭС в соответствии с описанием формальной структуры словосочетания на основе обобщенных синтагм, приведенных в нижней части табл.13.

***Шаг* 13.** Произвести преобразование текстового представления словарной статьи СЭС в его машинную форму. Входом в словарную статью служит обобщенная нормализованная синтагма словосочетания.

Алгоритм 5. Автоматическое формирование словаря СЭП

***Шаг* 1. Р**азделить текст на предложения и выполнить обработку каждого анализируемого предложения процедурой морфологического анализа.

***Шаг* 2.** Произвести упрощенный синтаксический анализ предложения: определить границы словосочетаний, главные и второстепенные члены предложения, построить дерево зависимостей предложения и сформировать «скелет» предложения.

***Шаг* 3.** Выделить в анализируемом предложении именные и глагольные словосочетания.

***Шаг* 4.** Для каждого словосочетания определить его позицию в предложении и длину, построить его обобщенную синтагму, установить главное слово и индекс его синтагмы.

***Шаг* 5.** Для главного слова словосочетания определить его позицию в словосочетании и установить его «внешнее» управление (его «хозяин» в предложении) и индекс синтагмы этого управляющего слова.

***Шаг* 6.** Внутри словосочетания определить все зависимые слова и их связи типа «хозяин-слуга», построить дерево зависимостей (главное слово представлено символом обобщенной синтагмы, для зависимых слов указывается номер позиции «хозяина»).

***Шаг* 7.** Для каждого слова словосочетания установить его однозначную грамматическую информацию (род, число, падеж, лицо).

***Шаг* 8.** Произвести автоматическую нормализацию исходного словосочетания.

***Шаг* 9.** По нормализованному представлению словосочетания построить его нормализованную синтагму.

***Шаг* 10.** Произвести поиск каждой сформированной поисковой синтагмы в словаре СЭС. В случае успешного поиска выполнить замену идентификатора словосочетания на идентификатор словаря СЭС. В случае неудачного поиска информацию о словосочетании поместить в дополнительный словарь и этому словосочетанию присвоить идентификатор по этому словарю.

Таблица 13

Результат преобразования текстового представления словосочетаний в его формализованное представление

n/n	*Идентификатор*	*Исходное словосочетание*	*Нормализованный «скелет» словосочетание*	*Позиция словосочетания в предложении*	*Тип словосочетания (N/V)*	*Кол-во слов*
1	*WC01*	*Трубопроводы высокого давления резервной котельной*	*трубопровод давление котельная*	*01*	*N*	*5*
2	*WC01*	*Трубопроводы высокого давления*	*трубопровод давление*	*01*	*N*	*3*
3	*WC02*	*резервной котельной*	*котельная*	*03*	*N*	*2*
4	*WC03*	*расположены на*	*располагать на*	*05*	*V*	*2*
5	*WC04*	*расположены от*	*располагать от*	*05*	*V*	*2*
6	*WC05*	*значительном расстоянии*	*расстояние*	*09*	*N*	*2*
7	*WC06*	*систем безопасности блока*	*система безопасность блок*	*11*	*N*	*3*
8	*WC07*	*безопасности блока*	*безопасность блок*	*12*	*N*	*2*

Результаты синтаксического анализа словосочетаний

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	***Пояснения к формальной структуре словосочетаний***
12345		*12*	*12*	12	12	123	12	Нумерация слов в словосочетании
NANAA	*NAN*	*AA*	*KF*	*KF*	AN	NNN	NN	Символов грамматических классов слов
АЦЙФФ	*АЦЙ*	*ФФ*	*ж7*	*жя*	*ФЙ*	*utF*	*tF*	Словосочетание в виде символов обобщенных синтагм (синтагмы расположены вертикально)
qfЦhh	*qfЦ*	*hh*	*qA*	*qA*	*iS*	*ASB*	*SB*	
N3151	*N31*	*2A*	*K1*	*K1*	2N	N12	N1	Дерево зависимостей словосочетания
13322	*133*	*22*	*00*	*00*	33	221	21	Информация о роде слов словосочетания
21111	*211*	*11*	*20*	*20*	11	211	11	Информация о числе слов словосочетания
12222	*122*	*22*	*20*	*20*	55	222	22	Информация о падеже слов словосочетания
00000	*000*	*00*	*00*	*00*	00	000	00	Информация о лице слов словосочетания
KN	*KN*	*NA*	*KF*	*KF*	NA	NN	NN	Тип внутренней синтаксической связи в словосочетании в символах классов слов
Aqжq	*Aqжq*	*AqФh*	*жq7A*	*жqяA*	*ЙSФi*	*uAtS*	*tSFB*	Тип внутренней синтаксической связи в словосочетании в символах обобщенной синтагмы
KN	*KN*	*NN*	*NK*	*NK*	*AN*	*NNN*	*NN*	Тип внешней управляющей синтаксической связи словосо-четания в символах классов слов
жqAq	*жqAq*	*AqЙЦ*	*AqЙЦ*	*Aqжq*	*7АЙS*	*яAuA*	*tSFB*	Тип внешней управляющей синтаксической связи словосочетания в символах обобщенной синтагмы

n/n	**Номер типа словосочетания**	**Информация о синтаксической и семантической структуре словосочетаний**
	1674	***жq7A-****жq-Aq-01-V-жq01-KF-00-20-00-00- «располагать на – расположены на»*
	1698	***жqяA****-жq-Aq-01-V-жq01-KF-00-21-020-00-« располагать от – расположены от»*
	0563	***AqЦfЙЦФhФh-****Aq-жq-01-N-Aq030101501-NANAA-13322-21111-12222-00000-«трубопровод давление котельная – Трубопроводы высокого давления резервной котельной»*
	0559	***AqЦfЙЦ-****Aq-жq-01-N-Aq0301-NAN-133-211-122-000-«трубопровод давление – Трубопроводы высокого давления»*
	1867	***ФhФh****-Фh-Aq-02-N-Aq030101501--AA-22-11-22-00-«котельная – резервная котельная»*
	1789	***ФiЙS****-ЙS-7A-02-N-02ЙS-AN-33-11-22-00-«расстояние – значительное расстояние»*
	1345	***uAtSFB****-uA-****яA-****01-N-uA0102-NNN-221-211-222-000-«систем безопасность блок – система безопасности блока»*
	1834	***tSFB-tS-uA-****01-N-tS01-NN-21-11-22-00-«безопасность блок – безопасности блока-»*

Таблица 14

Словарная статья словаря структур эталонных предложений

Идентификатор словаря СЭП	Информация о синтаксической и семантической структуре эталонного предложения
003486	***ЖqAqAЙSяAuuA=02****=«располагать-трубопровод–на-расстояние-от-система»=****03****=«Жq-2674#Aq-0563#ЙS-1754#uA-1248=****03****=020101030105=****04****=KNFNFN-PSFDFD-110302-*220102-*012222-000000-06=****06****=NWC-8=****08****=NSPO-2=****09****=AqЦfЙЦФhФh-жq7A-uAtSFB#01,05-06,02-08,02##AqЦfЙЦФhФh-жqяA-uAtSFB#01,05-06,02-08,02-11=****10****=«трубопровод–располагать-на-расстояние»=«трубопровод–располагать-от-система»=****11****=AqЦfЙЦФhФhжq7AФiЙSяAuAtSFB-NANAAKFANFNNN-13=****12****=«Трубопроводы высокого давления резервной котельной расположены на значительном расстоянии от систем безопасности блока»*

Таблица 15

Поисковая структура анализируемого предложения

01=ЖqAqAЙSяAuuA=02*=«располагать-трубопровод–на-расстояние-от-система»=****03****=жq-qЦfЙЦФhФh-7A-ЙSяA-ЙS-яA-uAtSFB=****04****=AqЦfЙЦФhФhжq7AФiЙSяAuAtSFB-13=****05****=«Трубопроводы высокого давления резервной котельной расположены на значительном расстоянии от систем безопасности блока»*

Блоки: 01-поисковая синтагма предложения (модифицированный «скелет» предложения); 02 – формализованное представление «скелета» предложение; 03 – последовательность структур словосочетаний в символах обобщенных синтагм; 04 – структура анализируемого предложения в символах обобщенных синтагм и длины предложения; 05 – анализируемое предложение в текстовом представлении.

***Шаг* 11.** По результатам обработки предложения построить словарную статью словаря СЭП в соответствии с описанием формальной структуры предложений на основе обобщенных синтагм, приведенных в табл.14.

***Шаг* 12.** Произвести преобразование текстового представления словарной статьи СЭП в его машинную форму. Входом в словарную статью служит обобщенная нормализованная синтагма предложения.

Теперь о том, как будет работать модернизированный алгоритм синтаксического анализа текстов на базе словаря СЭП. Основная идея этого алгоритма заключается в нахождении в словаре СЭП словарной статьи, содержащей тождественную или наиболее близкую анализируемому предложению синтаксическую структуру и соответствующее ему текстовое эталонное предложение с полным набором грамматической и семантической информации, ранее полученной на этапе традиционного синтаксического анализа. Здесь основным процессом является преобразование текстового вида анализируемого предложения в вид поисковой синтагмы, аналогичной по своей структуре словарной поисковой синтагме словаря СЭП. Такое преобразование можно осуществить упрощенной процедурой синтаксического анализа. Далее запускается механизм поиска на наибольшее совпадение левых частей поисковых синтагм анализируемого предложения и одной из синтагм словарных статей словаря. В случае полного совпадения поисковой синтагмы, расширенного представления синтаксической структуры и семантического наполнения эталонного предложения или полного совпадения синтаксической структуры и частичного совпадения с одной из словарных статей словаря, можно весь набор синтаксической информации приписать анализируемому предложению как результат его синтаксического анализа. В случае частичного совпадения синтаксических структур нужно выполнить дополнительный анализ тех частей поисковых синтагм анализируемого предложения, которые не совпали с частями словарной статьи словаря, с целью установления степени синтаксической близости соответствующим им частей предложения.

Автоматическое построение поисковой синтагмы предложения выполняется по тем же принципам, как и в случае формирования словарной статьи словаря СЭП. Отличие алгоритма построения поисковой синтагмы анализируемого предложения от алгоритма 5 в том, что в нем используются начальные шаги алгоритма, а результатом будет усеченная поисковая структура анализируемого предложения, приведенная в табл. 15.

ЗАКЛЮЧЕНИЕ

В настоящей статье описана модель синтаксической структуры текстов на основе обобщенных синтагм и показаны методы ее автоматического формирования. Предлагаемая синтаксическая модель базируется на машинной грамматике, в основу которой положена система флективных классов русских слов. Заложенное в теоретической концепции флективных классов слов русского языка жесткое соответствие между формой представления слов и их грамматической информацией позволило создать на

этой основе новые классы – *классы слов, имеющие одинаковые наборы грамматических признаков, соответствующие их формам представления в сходных контекстных окружениях.* При разработке этой синтаксической модели текстов мы исходили из следующей гипотезы: *одинаковым последовательностям обобщенных символов классов слов (обобщенным синтагмам) должны соответствовать одинаковые синтаксические структуры*. При этом предполагалось, что такая гипотеза верна для любых синтаксических моделей и может быть полезна при решении как глобальных, так и частных задач синтаксического анализа.

С помощью этой модели возможно реализовать новые решения ряда актуальных задач автоматической обработки текстовой информации в области информационного поиска и машинного перевода. Например, решение задачи автоматического синтаксического анализа текстов на основе предлагаемой модели позволит существенно повысить его быстродействие и качество и одновременно решить задачу его машинного обучения на основе корректировки результатов традиционного синтаксического анализа путем использования альтернативных методов назначения словам морфологических. синтаксических или семантических признаков. В рамках реализации этой модели были решены проблемы унификации синтаксических структур предложений и синтаксических конструкций, входящих в состав предложений.

В настоящее время авторы статьи проводят масштабные исследования по составлению словарей СЭС и словарей СЭП. Результаты этих исследований будут реализованы в новых версиях процедуры синтаксического анализа текстов, а также ряде других описанных выше процедур. Использование модели обобщенных синтагм в технологиях автоматической смысловой обработки текстовой информации позволит существенно повысить качество семантического анализа текстов и ускорить работу технологического процесса по обработке текстов на естественном языке.

СПИСОК ЛИТЕРАТУРЫ

1. Павлов Л.П. Информационный мир: истоки и новые реалии // Информатизация и связь. – 2012. – № 8. – С. 141-144.
2. Rabiner L., Juang B.-H. Fundamentals of Speech Recognition. – Prentice Hall, 1995. – 507 p.
3. Кипяткова И.С. Программно-алгоритмическое обеспечение создания синтаксическо-статистической модели русского языка по текстовому корпусу // Тр. СПИИРАН. – 2013. – № 24. – С. 332–348.
4. Vaičiūnas A. Statistical Language Models of Lithuanian and Their Application to Very Large Vocabulary Speech Recognition // Summary of Doctoral Dissertation. – Kaunas: Vytautas Magnus University, 2006. – 35 p.
5. Белоногов Г.Г., Зеленков Ю.Г., Новоселов А.П., Хорошилов Ал-др А., Хорошилов Ал-сей А. Метод аналогии в компьютерной лингвистике // Научно-техническая информация. Сер.2. – 2000. – № 1. – С. 21-31.
6. Аблов И.В., Козичев В.Н., Ширманов А.В., Хорошилов Ал-др А., Хорошилов А.А. Средства машинной грамматики русского языка (по Г.Г. Белоногову) // Научно-техническая информация. Сер.2. – 2018. – № 6. – С. 32-46; Ablov I. V., Kozichev V. N., Shirmanov A. V., Khoroshilov Al-dr A., Khoroshilov Al-ey A. The Tools of a Machine Grammar of the Russian Language (based on G.G. Belonogov) // Automatic Documentation and Mathematical Linguistics. – 2018. – Vol. 52, № 3. – P. 142-156.
7. Белоногов Г.Г., Калинин Ю.П., Хорошилов А.А. Компьютерная лингвистика и перспективные информационные технологии. Теория и практика построения систем автоматической обработки текстовой информации – М.: Русский мир, 2004. – 264 с.
8. Старовойтов А.В., Пошатаев О.Н., Прохоров С.Н., Хорошилов А.А. Методы автоматизированного составления и ведения словарей // Сб. Информатизация и связь / Центр информационных технологий и систем органов исполнительной власти. – 2013. – №3. – С. 91–97.

Материал поступил в редакцию 05.09.18.

Сведения об авторах

КАН Анна Владимировна – кандидат технических наук, начальник аналитического отдела Департамента координации и сопровождения Государственных программ ФГБУ "НИЦ "Институт имени Н.Е. Жуковского", Москва
e-mail: avkan@nrczh.ru

РЕВИНА Валерия Дмитриевна – студент Московского авиационного института (национальный исследовательский университет), Москва
e-mail: lerkovic96@ya.ru

РУСНАК Владислава Игоревна – студент Московского авиационного института (национальный исследовательский университет), Москва
e-mail: rusnak.vladislava.nn@yandex.ru

ХОРОШИЛОВ Александр Алексеевич – доктор технических наук, профессор МАИ, ведущий научный сотрудник Федерального исследовательского центра «Информация и управления» РАН, Москва
e-mail: khoroshilov@mail.ru

ХОРОШИЛОВ Алексей Александрович – кандидат технических наук, научный сотрудник Федерального исследовательского центра «Информация и управления» РАН, Москва
e-mail: a.a.horoshilov@mail.ru