Уфа : УГАТУ, 2008 Вестник УГАТУ • Управление, ВТ и И Т. 11, № 1 (28). С. 101–107

УДК 004:[81'42+81'23]

Н. К. КРИОНИ, А. Д. НИКИН, А. В. ФИЛИППОВА

АВТОМАТИЗИРОВАННАЯ СИСТЕМА АНАЛИЗА СЛОЖНОСТИ УЧЕБНЫХ ТЕКСТОВ

В статье приводятся этапы разработки и построения программного продукта для проведения исследований по анализу сложности параметров, влияющих на трудность восприятия учебных текстов. *Параметры текста*; *оценка сложности*; *трудность восприятия текста*

Учебные издания являются основным средством обучения в вузе. При этом учебный текст должен быть понятен любому студенту определенного уровня подготовленности со средним интеллектуальным развитием. «Понятен» — означает, что в результате прочтения текста в сознании студента должен сформироваться смысл, максимально близкий по содержанию авторскому замыслу. Степень близости характеризует трудность восприятия текста.

Трудность восприятия учебного текста определяется подготовленностью читателя (в перцептивном, когнитивном и аффективном аспектах) и свойствами текста. Свойства текста отражаются следующими основными характеристиками [13]:

1) связность (тематическая и логико-композиционная);

2) структурность (тематическая и логико-композиционная);

3) цельность;

4) функционально-смысловой тип (различные виды описательных и аргументативных текстов) [9];

5) информативность;

6) абстрактность изложения;

7) сложность лингвистических конструкций.

Характеристики (свойства) текста, от которых зависит трудность его понимания и усвоения, называются компонентами сложности текста. Все эти компоненты в совокупности называются сложностью текста [4].

Проблема трудности восприятия учебного текста может рассматриваться в двух аспектах: в аспекте генерации текста и в аспекте восприятия текста.

При генерации текста автор чаще всего неосознанно использует лексикон и лингвистические конструкции, усложняющие текст, чрезмерно насыщает учебный текст новой информацией. При объективной оценке сложности текста его автор на основании такой оценки может принять решение о целесообразности редактирования созданного текста с целью снижения его сложности.

Имея объективную оценку сложности учебного текста, преподаватель может оценить время, необходимое студенту на изучение текста, а студент, используя указания на элементы текста, определяющие его сложность, сможет при его изучении обратить особое внимание на данные фрагменты текста.

Таким образом, объективный анализ свойств учебного текста способствовал бы решению важной и актуальной задачи — улучшению его усвоения и, как следствие, повышению качества обучения.

Этой же задаче служит формирование перечня ключевых слов текста (точнее, ключевых понятий, включая понятия, обозначаемые в русском языке несколькими словами). Психолингвистическими исследованиями установлено, что перечень ключевых слов, будучи предъявлен читателю перед чтением текста, способствует лучшему усвоению текста. А. А. *Залевская* объясняет это тем, что перечень ключевых слов играет роль «смысловых вех», позволяя читателю еще до восприятия текста спрогнозировать его содержание. При этом «авторское» включение тех или иных слов в перечень ключевых носит субъективный характер. При таком анализе автор склонен учитывать и имплицитную информацию, которая в явном виде в тексте отсутствует. Поэтому важно формировать пере-

Контактная информация: (347)272-80-45

чень ключевых слов по объективным критериям.

Авторами была выдвинута гипотеза о возможности формализации поиска в русскоязычном тексте определений понятий, количество которых повышает информативность текста, абстрактных понятий и ряда лингвистических конструкций, усложняющих текст. Разработанные алгоритмы формализации поиска были реализованы в созданном программном продукте.

Алгоритмы автоматического формирования перечня ключевых слов, основанные на частотном анализе текста, известны [11].

НАЗНАЧЕНИЕ РАЗРАБАТЫВАЕМОГО ПРОГРАММНОГО ПРОДУКТА

Создаваемый ПП должен обеспечить выполнение следующих функций:

1) автоматическое выделение ключевых слов;

2) автоматический поиск параметров сложности текстов и их подсчет;

3) осуществление диалога с пользователем с целью проверки правильности выделенных параметров;

4) сохранение результатов анализа сложности текста и поиска в нем ключевых слов в отдельных файлах;

5) реализация возможности работы с документами в форматах doc, txt, rtf.

СТРУКТУРА ПРОГРАММЫ

Обобщенная структура программы представлена на рис. 1.

АЛГОРИТМЫ РАБОТЫ ПРОГРАММЫ

Алгоритм выделения ключевых слов и описание соответствующих диагностирующих признаков приведены в статье [7].

Для разработки алгоритма поиска и подсчета параметров сложности, влияющих на трудность восприятия текста, были сформированы таблицы диагностирующих признаков для каждого из параметров сложности.

ДИАГНОСТИРУЮЩИЕ ПРИЗНАКИ ОПРЕДЕЛЕНИЯ ПОНЯТИЯ

Информативность текста характеризуется количеством семантической информации, количеством новых для читателей сведений [15]. Форма представления в тексте таких сведений (функционально-смысловой тип текста) может быть различной: определение, описание, повествование, доказательство и т. п.

Информативность текста в программе оценивалась количеством приводимых в тексте определений новых понятий. Для учебного текста характерным является введение новых понятий, направленных на расширение картины мира обучающихся. В большинстве случаев понятие вводится классической формой логического определения [6]. Для данной формы характерен ряд специфических лингвистических конструкций. При анализе текста элементы этих конструкций могут выступать в качестве признаков, позволяющих диагностировать определение [12]. Признаки лингвистических конструкций логического определения, реализованные в алгоритме программы, представлены в табл. 1.

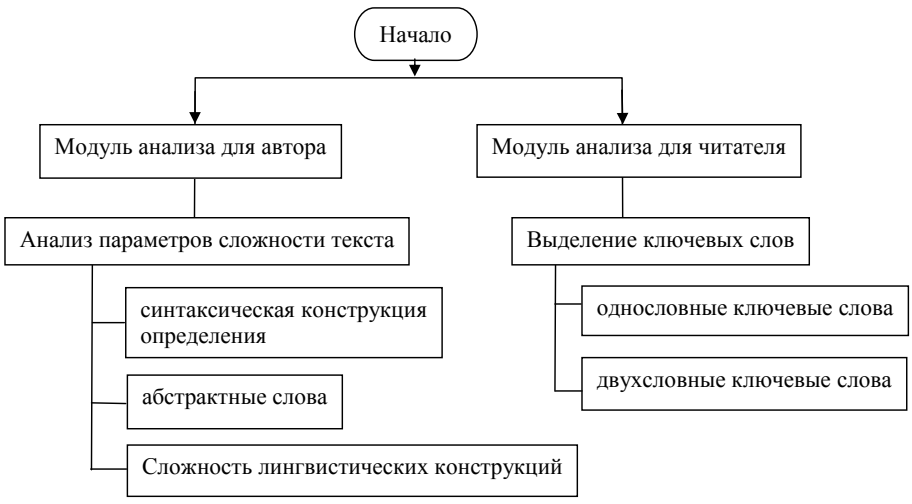


Рис. 1. Обобщенная структура программы

Алгоритм поиска конструкции логического определения предполагает поиск в тексте любых из перечисленных признаков и их подсчет.

Программа, реализующая алгоритм, составленный на основе признаков, приведенных в табл. 1, не позволяет находить определения, данные в неявной форме. Работу с программой предполагается проводить в автоматизированном режиме, автоматически маркируя найденные фрагменты текста, похожие на определения понятий, и оставляя возможность принятия окончательного решения за экспертом, так как возможны варианты, когда найденная лингвистическая конструкция, соответствующая признакам, в действительности не является определением. Разработанная программа достаточно полно производит поиск определений: по результатам тестирования на корпусе текстов объемом 15 000 знаков с учетом пробелов, программа выделила 90% определений, заданных в явной форме, остальные 10% — это фрагменты, которые соответствовали лингвистической конструкции логического определения, но по смыслу не являлись таковыми, или в тексте присутствовали определения, признаки которых в диагностирующей таблице не описаны.

Таблица 1

Диагностирующие признаки лингвистической конструкции логического определения

№ п.п.	Признаки
1	Слово–пробел–тире–пробел–Это
2	Слово–пробел–тире–пробел–слово[1]
3	слово–пробел–«есть»–пробел–слово
4	называ . . .
5	понима . . .
6	представляя . . . собой
7	означа . . .
8	обознача . . .
9	определе . . .
10	считае . . .

ДИАГНОСТИРУЮЩИЕ ПРИЗНАКИ ХАРАКТЕРИСТИКИ «АБСТРАКТНОСТЬ ИЗЛОЖЕНИЯ»

Кроме информативности на сложность понимания текста, по мнению ряда исследователей [4], влияют абстрактность изложения содержания и сложность лингвистических конструкций текста.

Для оценки абстрактности изложения обычно принимается доля слов в составе текста, обозначающих абстрактные смысловые объекты, то есть объекты, не доступные непосредственному чувственному восприятию, например, открытие, величина и т. д. *Я. А. Микк* [4] полагал, что чем больше доля в тексте абстрактных слов, тем сложнее текст. Он предложил использовать в качестве диагностических признаков для отнесения слов текста к категории абстрактных слов абстрактные суффиксы:

1) -й- (ье), -ти(е), -ни(е), -ени(е), -ани(е), -и(е). Простой суффикс образует названия отвлеченных понятий и состояний:

здоровье, веселье, бессмертие, многословие, великодушие, открытие, нажатие;

2) -ств(о), -еств(о), -честв, -овств, -ничеств, -тельств. Посредством этих суффиксов образуются существительные от существительных, прилагательных, глаголов [3]:

учительство, невежество, упорство, упрямство, коварство;

3) -аци(я). Суффикс создает значение осуществления известного мероприятия, процесса:

мобилизация, организация, реализация, активизация;

4) -ость(-есть), -емость,-ность, -емость, -имость. Суффикс обозначает отвлеченное понятие о качестве:

храбрость, резвость, близость, опрометчивость, успеваемость, рождаемость, воспитанность;

5) -изм, суффикс служит для обозначения общественно-политических и научных течений и учений, а также отвлеченных качеств и обычно употребляется с основами заимствованных слов:

капитализм, коммунизм, комизм, историзм;

6) -изн(а), суффикс создает обозначения отвлеченных качеств:

кривизна, прямизна;

7) -от(а), суффикс используется для указания на обозначенный по качеству предмет:

мокрота, прямота, чернота, правота, простота [2];

8) -ин(а), суффикс также обозначает отвлеченные качества:

толщина, седина, величина, старина, ширина;

9) -щин(а), -овщин(а) — сложные суффиксы, образовавшиеся из суффиксов прилагательных -ск и -ин, обозначают общественные явления с резко отрицательной оценкой. Слова с суффиксом -щин в большинстве случаев образуются от имен лиц, деятельность

которых служит характерным признаком эпохи, режима, общественных явлений [2]:

иностранщина, хлестаковщина, чертовщина, обломовщина;

10) -к(а) — суффикс, обозначающий конкретные действия или производственный процесс [2].

В алгоритме программы данный суффикс не рассматривается, так часто слова с таким суффиксом получают значения конкретных и вещественных существительных, иногда совмещая двоякое значение, например, замазка окон и густая замазка, постройка завода и заводские постройки [3];

11) -н(я), -отн(я). В алгоритме программы данный суффикс не рассматривается, так как не всегда слово с таким суффиксом может являться абстрактным:

возня, грызня, брехня;

12) -ик(а), -истик(а), -астик(а). Суффикс служит для обозначения научной дисциплины, отдела науки, искусства, сферы деятельности:

грамматика, евгеника, фонетика, оптика [2];

13) -тив(а), -атив(а) — суффикс, выделяемый в заимствованных научно-книжных словах, преимущественно с отвлеченными значениями: прерогатива, инициатива, инвектива [2];

14) -ур(а), -юр(а), данный суффикс встречается в названиях областей культуры и искусства, например, брошюра, миниатюра, в названии действия, например, процедура, мануфактура, в названии профессии — профессура, адвокатура и т. д. [2]. В алгоритме данный суффикс не рассматривается.

В алгоритме программы рассматриваются только наиболее часто употребляемые суффиксы отвлеченных понятий, представленные выше.

Созданная программа идентифицирует и маркирует цветовым фоном слова, содержащие абстрактные суффиксы. Лицо, принимающее решение, в диалоговом режиме может принять окончательное решение об отнесении маркированного слова к категории абстрактных слов [10]. Программа достаточно полно производит поиск абстрактных слов. По результатам тестирования, программа на 97% точно выявила абстрактные слова, что позволяет судить о достоверности результатов программы.

ДИАГНОСТИРУЮЩИЕ ПРИЗНАКИ ХАРАКТЕРИСТИКИ «СЛОЖНОСТЬ ЛИНГВИСТИЧЕСКИХ КОНСТРУКЦИЙ»

Сложность лингвистической конструкции текста характеризуется:

- количеством длинных слов в тексте (слов с тремя и более слогами);
- количеством (долей) предложений текста, содержащих длинные слова;
- средней длиной слова в тексте [8];
- средней длиной предложения, измеряемой количеством слов, входящих в него [8];
- количеством в предложениях текста причастий и деепричастий;
- количеством (долей) предложений текста, содержащих причастия и деепричастия;
- количеством (долей) сложных предложений текста.

Полагается, что чем больше сложных лингвистических конструкций в тексте, тем текст сложнее для восприятия [14].

В то же время авторы полагают, что наличие определенного процента сложных лингвистических конструкций должно быть обязательным условием составления учебного текста, так как сложные лингвистические конструкции развивают логику учащегося и повышают уровень его интеллекта.

В настоящее время в программе реализован модуль поиска сложносочиненных предложений. Диагностирующими признаками, позволяющими отнести предложение к категории сложносочиненных, является наличие в предложении простых или сложных сочинительных союзов (табл. 2).

Таблица 2

Диагностирующие признаки сложносочиненных предложений

Сочинительные союзы (простые)	Сочинительные союзы (сложные)
и, а, но , да, или, тоже, также, либо	ни – ни то – то как – так не только – но и не то – не то

Например, «*Высоко в небе сияло солнце, а горы зноем дышали в небо*» (М. Горький) [1], два простых предложения (1) высоко в небе сияло солнце; 2) горы зноем дышали в небо) связаны союзом «а».

«*Сквозь серый камень вода сочилась, и было душно в ущелье тёмном, и пахло гнилью*» (М. Горький) [1]. Три простых предложения связаны союзом «и».

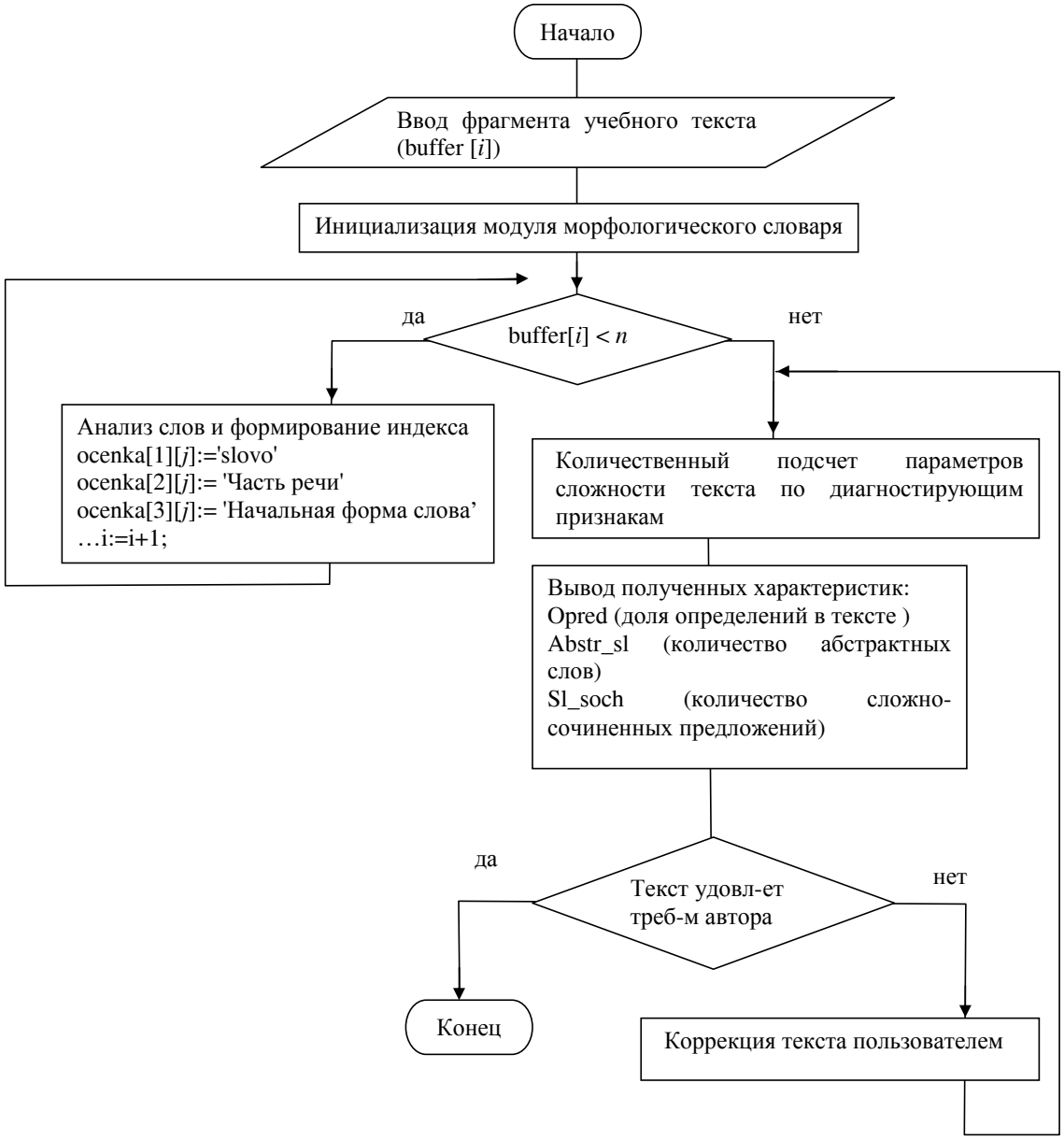


Рис. 2. Структура модуля анализа для автора

Созданная программа идентифицирует и маркирует цветовым фоном сложносочиненные предложения. Разработанная программа достаточно полно производит поиск сложносочиненных предложений. По результатам тестирования программы, правильно было выявлено 96% сложносочиненных предложений, остальные 4% невыявленных предложений имели сложную конструкцию, не предусмотренную авторами в диагностирующей таблице.

По выделенным диагностическим признакам параметров сложности текста была разработана программа, подсчитывающая количественные характеристики рассматриваемых параметров.

РАЗРАБОТКА ПРОГРАММНОГО ПРОДУКТА

В качестве инструментального средства проектирования (пакета прикладных программ) был выбран продукт Borland Delphi 6.0. Этот программный пакет, разработанный фирмой Borland International, общепризнанный лидер среди инструментов для создания приложений и систем, функционирующих на платформе Windows.

В программе реализована функция работы не только с текстовыми документами в формате .txt, но и документами Word в формате .doc, что не всегда доступно в некоторых программах, работающих с текстами.

Разрабатываемая программа предполагает переработку и дальнейший анализ текста, на-

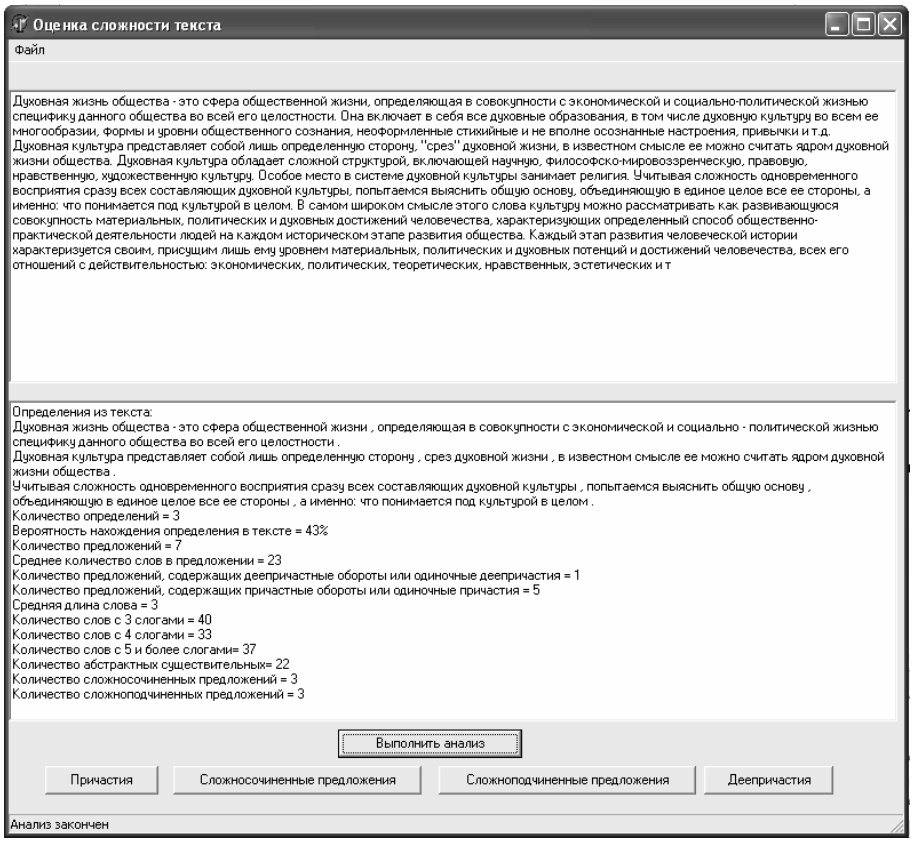


Рис. 3. Окно модуля анализа для автора

писанного на естественном языке. Для работы с естественным языком были использован морфологический словарь [5].

Морфологический словарь содержит все словоформы русского языка. Морфологический словарь для русского языка (также называемого «Диалинг») базируется на грамматическом словаре А. А. Зализняка 1987 г. и включает на данный момент 161 тысячу слов.

В программе анализ наличия параметров сложности в тексте и их поиск производился на основе данных о словоформе каждого отдельного слова из текста.

В качестве метода представления и сбора данных о слове был выбран метод индекса, известный также как метод предметного указателя, когда для каждого термина записывается набор страниц, где может этот термин находиться. Аналогичный принцип использовался для работы программы, был создан многомерный массив, хранящий набор информации о слове. В набор информации о слове вошли:

- исходная форма слова;
- морфологическая часть речи;
- начальная форма слова (словоформа);
- количество гласных;
- номер нахождения в тексте.

Анализ многомерного массива по заложенным диагностирующим таблицам позволяет выявить наличие параметров сложности в тексте, выполнить их подсчет и вывести данные на экран. По результатам первичного анализа выводится таблица количественных характеристик параметров сложности текста, далее автору предлагается отдельно поработать с каждым параметром текста, а именно — просмотреть каждую группу найденных параметров сложности текста. Автор текста может либо оставить в тексте все как есть, либо упростить конструкцию предложения, убрать абстрактное слово, изменить определение и провести повторный анализ.

Структура модуля анализа для автора, производящая подсчет количественных характеристик параметров сложности, представлена на рис. 2.

На рис. 3 приведен вид окна модуля анализа для автора после выполнения первичного анализа текста.

ВЫВОДЫ

1. Разработан программный продукт — инструмент для автоматического поиска параметров сложности текста, их подсчета и автоматизированного анализа;

2. Разработанный «Модуль анализа для читателя» позволяет автоматически выделять ключевые слова;

3. Разработанный «Модуль анализа для автора» позволяет автору редактировать текст с учетом найденных параметров;

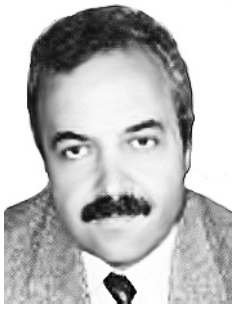
4. Программа может выполнять сохранение и накопление данных в отдельных файлах;

5. В программе реализована возможность работы с документами в форматах doc, txt, rtf.

СПИСОК ЛИТЕРАТУРЫ

1. **Бабайцева, В. В.** Русский язык : теория / В. В. Бабайцева, Л. Д. Чеснокова. М. : Просвещение, 1994. 256 с.
2. **Виноградов, В. В.** Русский язык (грамматическое учение о слове) / В. В. Виноградов. М. : Гос. учеб.-пед. изд-во м-ва просвещения РСФСР, 1947. 731 с.
3. **Гвоздев, А. Н.** Современный русский литературный язык : Ч. 1 / А. Н. Гвоздев. М. : Гос. учеб.-пед. изд-во м-ва просвещения РСФСР, 1958. 421 с.
4. **Микк, Я. А.** Оптимизация сложности учебного текста : в помощь авторам и редакторам / Я. А. Микк. М. : Просвещение, 1981. 119 с.
5. **Морфологический** словарь [Электронный ресурс] (http://www.aot.ru).
6. **Новиков, А. И.** Семантика текста и ее формализация / А. И. Новиков / АН СССР. Ин-т языкознания. М. : Наука, 1983. 97 с.
7. **Никин, А. Д.** Информационная система анализа учебного текста / А. Д. Никин, Н. К. Криони, А. В. Филиппова // Телематика'2007 : тр. XIV Всерос. науч.-метод. конф. Т. 2. ГосНИИ информ. технол. и телекоммуникаций «Информика», 2007. С. 463–465.
8. **Оборнева, И. В.** Автоматизированная оценка сложности учебных текстов на основе статистических параметров : дис.... канд. пед. наук / И. В. Оборнева. М., 2006, 120 с.
9. **Буре, Н. А.** Основы научной речи / Н. А. Буре [и др.]. СПб. : Филологический факультет СПбГУ; М. : Академия, 2003. 272 с.
10. **Орлов, А. В.** Экспертные оценки : учеб. пособие / А. В. Орлов [Электронный ресурс] (http://www.aup.ru/books/).
11. **Попов, А.** Поиск в Интернете — внутри и снаружи. Эффективная методика поиска информации в сети Интернет / А. Попов [Электронный ресурс] (http://www.dist-cons.ru/modules/searchinf/z/a14/inter.net.ru/7/18.html).
12. **Пушкина, Е. С.** Теоретико-экспериментальное исследование структурно-семантических параметров текста : автореф. дис.... на соиск. уч. ст. канд. филолог. наук. Кемерово, 2004. 25 с.
13. **Редченкова, Г. Д.** «Учительская экспертиза» учебника (критерии, по которым может быть оценен учебник) / Г. Д. Редченкова [Электронный ресурс] (http://www.iro.yar.ru:8101/).
14. **Райс, К.** Классификация текстов и методы перевода / К. Райс [Электронный ресурс] (http://www.philology.ru/).
15. **Сохор, А. М.** Логическая структура учебного материала / А. М. Сохор. М. : Педагогика, 1974. 119 с.

ОБ АВТОРАХ



Криони Николай Константинович, проф., проректор. УГАТУ. Дипл. инж-мех. по техн. машиностр. (УАИ, 1971). Д-р техн. наук по трению и износу в машинах (РГУПиГ им. И. М. Губкина, 1985). Иссл. в обл. трибологии контактн. взаимодействия, методики и организации учеб. процесса в высш. школе.

Никин Алексей Дмитриевич, доц., нач. отдела образ. технол. Дипл. инж.-электромех. по автоматизации, комплексн. механизации машиностр. (УАИ, 1978). Канд. техн. наук по автоматизации технологич. процессов и производств (УГАТУ, 1999). Иссл. в обл. автоматизации технологич. процессов, методики и организации учеб. процесса в высш. школе.

Филиппова Анастасия Владимировна, асп. каф. АСУ. Дипл. экон.-матем. по матем. методам в экономике (УГАТУ, 2006). Иссл. в обл. управл. в соц. и эконом. системах.