Исправление ошибок в русскоязычных текстах

Проблема речевых ошибок

Использование естественного языка в качестве средства общения (***речевая деятельность*** человека) неизбежно сопровождается теми или иными нарушениями языковых правил. Такие нарушения - вне зависимости от того, обусловлены они неполнотой знаний человека о языке или же случайными сенсомоторными «сбоями» (описки, опечатки, оговорки) - мы будем называть ***речевыми ошибками***.

В идеале обработка речевой ошибки предполагает соотнесение ошибочной речевой единицы с полным описанием языка и с контекстом рассматриваемого коммуникативного процесса. Лингвист (или другой специалист), занимающийся исследованием каких-либо теоретических аспектов проблемы речевых ошибок, например, их классификацией, и располагающий источниками, в которых содержится исчерпывающее описание единиц и правил того или иного естественного языка (словари, своды правил), находится в ситуации, достаточно близкой к такому идеалу.

В случае же повседневной речевой практики – непосредственного (диалог) или опосредованного (чтение текста) речевого взаимодействия рядовых носителей языка - ситуация иная. Лингвистические знания рядового носителя языка неполны, воспользоваться справочной литературой он может далеко не всегда, а сам факт ошибки никаким явным образом в анализируемом тексте не указан.

Обнаружить речевую ошибку в этой ситуации непросто. Действительно, для получателя сообщения (реципиента) внешним признаком речевой ошибки служит появление в тексте какой-либо незнакомой ему речевой единицы. Однако такая «подозреваемая» речевая единица может оказаться и правильной конструкцией или формой (например, просторечным вариантом или термином), не знакомой реципиенту.

С другой стороны, абсолютно правильная на первый взгляд единица может быть ошибкой, обнаружить которую удается лишь на «высших» этапах анализа. Так, в предложении *«Пуск ракеты осуществляется нажатием краской кнопки»* все слова известны, синтаксические связи правильны; опечатка обнаруживается только на семантическом/ смысловом уровне.

Если одним из участников общения является компьютерная система, положение становится еще более сложным. И лингвистические знания, и интеллектуальные способности (в том числе - в плане работы с языком) такого «собеседника» пока весьма скромны. Однако системы обнаружения и исправления ошибок применяются достаточно широко и успешно.

Отметим еще одно обстоятельство. Как бы ни разнились характер использования и назначение АОТ-систем (системы машинного перевода, автоматического реферирования или индексирования, работающие в пакетном режиме; системы обеспечения диалога с машиной на естественном языке), оснащение их средствами обнаружения и исправления речевых ошибок повышает устойчивость и эффективность функционирования таких систем, облегчает (в случае диалоговых систем) процесс общения человека с ЭВМ.

Классификация речевых ошибок

Первый критерий классификации речевых ошибок, в соответствии с которым ошибки подразделяются на мотивированные и случайные, связан с понятием индивидуальной языковой модели. ***Индивидуальная языковая модель*** (ИЯМ) - это то подмножество языковых единиц и правил, которое усвоил и использует в своей речевой практике конкретный носитель некоторого естественного языка. Субъективное преломление языка (как знаковой системы социального уровня) в процессе его усвоения приводит к тому, что в ИЯМ не попадают (или попадают в искаженном варианте) некоторые языковые единицы и правила языка.

Поэтому в речи конкретных носителей языка начинают проявляться некоторые индивидуальные особенности, вступающие в противоречие с языковыми нормами или нет.

В первом случае мы имеем дело с мотивированными речевыми ошибками - точнее, с ошибками, мотивированными особенностями ИЯМ конкретного носителя языка (автора анализируемого системой текста). К ошибкам такого рода относятся, например, ошибки в словоизменении (*контейнер**а́*** - в форме именительного падежа множественного числа), орфографические ошибки в основах (*ед**е**ница*), некоторые пунктуационные ошибки, смешение слов-паронимов (*представить* - *предоставить*), нарушение лексической сочетаемости (*делать горе*), искажение фразеологизмов (*не так страшен черт, как его малютки*).

Ошибки, обусловленные внешними по отношению к ИЯМ факторами: сбой речевого аппарата человека, несвоевременное переключение регистра клавиатуры, нажатие соседней клавиши, сбой на линии связи с ЭВМ - мы будем называть случайными.

Как правило, мотивированные речевые ошибки регулярно повторяются в речи носителя языка, а случайные ошибки могут как повторяться (например, при западании клавиши), так и не повторяться. Отметим, что иногда отличить случайную ошибку от мотивированной сложно. Так, употребление слова *представить* вместо *предоставить* в контексте *представлено право* может быть или результатом случайной ошибки (пропуск буквы), или результатом мотивированной ошибки (смешения паронимов).

Мотивированные речевые ошибки различаются степенью серьезности (грамматичности). Помимо серьезных, абсолютно недопустимых грамматических ошибок - типа орфографических ошибок в основах или смешения слов - рассматриваются и ошибки, в результате которых появляются «полуграмматичные» формы (*контейнерá*, *сидевши*), которые имеют в словарях стилистические пометы: просторечное, устарелое, разговорное, областное и др.

Следующий критерий классификации ошибок (мотивированных и случайных) связан с языковыми уровнями, нормы (правила) которых оказываются нарушенными в результате речевых ошибок. В соответствии с этим критерием речевые ошибки можно классифицировать следующим образом:

1) орфографические ошибки: пропуск одной буквы, замена одной буквы, перестановка двух рядом стоящих букв, одна лишняя буква (отдельно может рассматриваться случай удвоения буквы), замена буквы русского алфавита буквой латиницы и др.;

2) морфологические (словоизменительный уровень) ошибки: ошибки в окончаниях (флексиях) при склонении и спряжении слов (рассматриваются различные подклассы таких ошибок), употребление отсутствующих в языке форм слов, несоблюдение правил чередования в основе, употребление незнакомых АОТ-системе вариантов слов, испытывающих колебания в роде, одушевленности;

3) синтаксические ошибки: ошибки в моделях управления слов-предикатов, пунктуационные ошибки, нарушение нормативного порядка слов (в том числе - в устойчивых словосочетаниях), вставка пробела внутрь слова, пропуск пробела (отдельно могут рассматриваются случаи слитного и раздельного написания частиц *не* и *ни*);

4) лексико-семантические ошибки: употребление слов в неверном значении, нарушение лексической сочетаемости, семантические противоречия.

Диагностика речевых ошибок

Методы обнаружения и исправления орфографических и морфологических ошибок в текстах широкой тематики базируются на представлении о тексте как о цепочке независимо появляющихся словоформ. Известно три основных метода обнаружения орфографических ошибок - статистический, полиграммный и словарный.

При статистическом методе словоформы, обнаруживаемые в тексте, упорядочиваются согласно частоте их встречаемости. Искаженные слова оказываются среди малоупотребительных слов в конце списка.

При полиграммном методе все встречающиеся в тексте двух- или трёхбуквенные сочетания (полиграммы) проверяются по таблицам, содержащим информацию об их допустимости в русском языке. Если в словоформе имеются недопустимые полиграммы, то она считается неправильной.

При словарном методе все входящие в текст словоформы проверяются по компьютерному словарю. Если словарь такую форму допускает, она считается правильной, а иначе либо сразу признаётся ошибочной, либо предъявляется человеку.

В настоящее время первые два метода практически не используются, т.к. уже есть хорошие компьютерные словари, достаточно большие по объёму и с эффективным доступом.

Диагностика же и исправление синтаксических, пунктуационных и лексико-семантических ошибок предполагает взгляд на текст как на последовательность связанных единиц, комбинирование которых имеет свои закономерности. Подходы к автоматизации выявления и коррекции этих ошибок можно разбить на две группы: синтаксически-ориентированные подходы и подходы, основанные на концептуальных фреймах. Последние больше пригодны для систем, работающих в строго ограниченных предметных областях. Для текстов широкой тематики предназначены синтаксически ориентированные подходы. Сначала поступившее на

вход предложение обрабатывается средствами грамматики, рассчитанной на синтаксически правильный текст. Если такая проверка обнаруживает дефекты синтаксической структуры, некоторые условия ослабляются. Какие грамматические правила смягчаются, зависит от учитываемых системой ошибок. Например, в русских текстах иногда оказывается пропущенной запятая, обособляющая причастный оборот в постпозиции. Для того, чтобы такое предложение могло быть обработано, требуется временная отмена условия (присутствующего в каноническом правиле) обязательного наличия запятой. Однако ослабление канонических правил неизбежно влечёт за собой возрастание числа возможных интерпретаций. При этом нельзя опознать ошибочный текст прежде, чем будет закончен анализ средствами канонической грамматики. Другой подход предлагает сначала использовать слабую грамматику, а затем подвергнуть обрабатываемое предложение фильтрации на основе строгих требований правильности. Но при этом наличие ошибки предполагается более вероятным, чем соблюдение норм грамматики.

Также отметим, что описанные методы позволяют автоматически обнаружить ошибку только тогда, когда не удаётся построить связный синтаксический граф для рассматриваемого предложения. Однако ошибки, при которых возможно получение формально приемлемой, но по сути неверной интерпретации, остаются невыявленными. При этом никаких сообщений об ошибках не поступает.

Система комплексного контроля качества текста ЛИНАР

Функции системы ЛИНАР; сценарии работы с системой

Построение автокорректоров сталкивается с рядом принципиальных и не решенных пока в полном объеме проблем: компактное хранение словарей, эффективные методы морфологического и синтаксического анализа и т.д. Тем не менее на очереди - создание систем, способных производить более сложное по сравнению с автокорректорами автоматическое или автоматизированное редактирование текстов на естественном языке. В идеале же необходима система, выполняющая функции научного редактора - человека, осуществляющего литературную и научную правку научно-технических текстов. Такое направление развития представляет разрабатывавшаяся в 1986-1990 гг. на кафедре алгоритмических языков факультета ВМК МГУ система ЛИНАР (ЛИтературно-НАучный Редактор) - интеллектуальная система комплексного контроля качества и редактирования русскоязычных текстов.

Суть подхода заключалась в существенном расширении возможностей имевшихся в то время автокорректоров за счет:

- ограничения предметной области, к которой относились обрабатываемые тексты (методы, алгоритмы и программы обработки данных телеметрии на многопроцессорных вычислительных комплексах);
- ограничения видов текстов (научно-технические отчеты, деловая переписка);
- использования средств синтаксического и семантического анализа текста;
- привлечения более полных моделей русского языка.

Пользователем ЛИНАР является человек, оценивающий с помощью системы качество некоторого текста с позиций лица, которому адресован этот текст (адресата), и вносящий в текст необходимые исправления. В качестве адресата могут выступать литературный или научный редактор, корректор, потенциальные читатели (конструкторы, программисты, руководители). Пользователем ЛИНАР может быть, например, автор обрабатываемого текста, желающий взглянуть на него «со стороны», или научный руководитель работы, обеспокоенный терминологическими и стилистическими неувязками в текстах разделов, подготовленных различными участниками проекта.

Обработка текста с помощью системы ЛИНАР включает в себя в общем случае несколько циклов (как и при подготовке текста «вручную»), каждый из которых оформляется как самостоятельный сеанс работы с системой. В начале сеанса пользователь формирует задание на обработку текста, для выполнения которого система загружает необходимые информационные модули и вызывает программы контроля текста. Каждая программа проверяет некоторое определенное свойство текста, т.е. реализует одноаспектный контроль текста. Таким образом, в структурном плане систему ЛИНАР можно считать пакетом прикладных программ; сеанс работы с ней состоит из серии одноаспектных проверок текста или его фрагментов.

Основная технологическая схема использования системы ЛИНАР предусматривает, что текст хранится на машинных носителях и обрабатывается программами контроля, формирующими

протокол замечаний по тексту (иногда система предлагает свой вариант исправления). Далее пользователь просматривает эти замечания и, если он с ними соглашается, вносит необходимые изменения в текст с помощью текстового редактора. Измененная версия текста может быть объектом обработки в следующем сеансе. В зависимости от объема текста пользователь может выбрать диалоговый или пакетный режим работы с системой. В последнем случае протокол замечаний формируется на внешнем носителе.

Отметим, что используемые в ЛИНАР знания позволяют системе фиксировать различные типы конфликтных ситуаций (и формировать соответствующие замечания). Однако как бы полны ни были знания ЛИНАР, обнаружить все неточности, противоречия, неопределенности система самостоятельно не может. Поэтому часть программ контроля собирает некоторую вспомогательную информацию о тех или иных характеристиках (свойствах) текста, не давая ей оценки.

Например, при написании отдельных фрагментов текста разными авторами для обозначения одной и той же сущности могут быть использованы различные термины, что усложняет понимание текста. Автоматическое обнаружение подобных конфликтов требует привлечения глубоких знаний о понятийном и терминологическом аппарате предметной области, и в ЛИНАР не реализуется. Однако в составе системы имеется программа контроля, которая может сформировать по фрагментам текста списки используемых терминологических словосочетаний. На основе этой информации решить терминологические проблемы человеку будет значительно проще, чем при обработке текста «вручную».

ЛИНАР не только обнаруживает неточности, ошибки, но и может «объяснить» пользователю суть своих замечаний, а также предложить способы устранения ошибок. Так, например, в случае орфографической ошибки система предлагает свой вариант исправления слова, в случае нарушения естественного порядка слов - правильный порядок слов и т.д. Рекомендации системы призваны помочь пользователю в улучшении текста, направляют его деятельность.

База знаний системы

Контроль текста, осуществляемый системой ЛИНАР, основывается на использовании знаний о том, что такое правильный, хороший текст. Совокупность этих знаний называется контролирующими знаниями, или К-знаниями. При формировании К-знаний учитывались результаты лингвистических, психологических работ, исследований по эргономике; принят во внимание опыт редакторов, корректоров, нормоконтролеров.

К-знания должны обеспечить возможность оценки текста с различных сторон:

- соответствие общеязыковым нормам;
- соответствие «внешним» нормам, например, требованиям ГОСТов, регламентирующих форму изложения материала в научно-технических документах;
- сложность восприятия текста потенциальным читателем;
- семантическая корректность текста (соответствие выявляемых в тексте семантических отношений и понятийной модели предметной области).

Часть К-знаний (процедурная составляющая) представлена программами одноаспектного контроля. Каждая программа фиксирует строго определенное свойство текста или строго определенный дефект текста (конфликтную ситуацию). Затем формируется соответствующее диагностическое сообщение, которое, в зависимости от выбранного режима работы, либо сразу предъявляется пользователю, либо включается в протокол замечаний.

Важным компонентом информационного обеспечения системы ЛИНАР является и лингвистическая база знаний, содержащая базовые общие знания о русском языке. Кроме того, ЛИНАР использует тематический словарь и тезаурус предметной области, к которой относятся обрабатываемые тексты, и описания нормативных требований, предъявляемых к текстам. Соответствующие информационные массивы создавались разработчиками системы на основе общеязыковых и предметно-ориентированных словарей и справочников, Государственных стандартов и отраслевых инструкций по оформлению текстовых документов.

База знаний ЛИНАР содержит также заранее формируемый - и пополняемый в ходе эксплуатации системы - **банк адресатов**: конкретных читателей или определенных однородных групп читателей (конкретный руководитель научно-исследовательского проекта; конкретный представитель руководства организации-заказчика; инженеры, которые будут создавать описываемый программно-аппаратный комплекс и др.). Настройка на адресата производится в начале очередного сеанса работы с ЛИНАР. При такой настройке могут меняться базовые и

тематические лингвистические знания (состав словаря, совокупность грамматических правил), степень жесткости требований по соблюдению тех или иных норм и условий.

Чтобы задать эту информацию, следует указать имя одного из известных ЛИНАР адресатов (или идентификатор известной группы адресатов) и выбрать значения дополнительных параметров программ контроля.

С помощью такой настройки удается моделировать процесс восприятия текста разными адресатами и, следовательно, оценивать качество текста с разных точек зрения.

Таким образом, К-знания ЛИНАР (которые служат критерием корректности текста и используются для обнаружения «дефектов» текста - отклонений от требований, предъявляемых К-знаниями) формируются динамически в каждом конкретном сеансе работы с системой и являются комплексными по своей природе. Они включают как процедурные знания об исследуемом аспекте текста (воплощенные в соответствующих программах контроля), так и декларативные знания, фильтруемые и конкретизируемые в начале каждого сеанса.

Обнаруженные программой контроля несоответствия текста и К-знаний могут быть устранены двумя способами:

- путем внесения изменений в текст (это наиболее частый случай: несоответствие - суть ошибка, допущенная в тексте, которую необходимо исправить);
- путем изменения К-знаний системы.

Заметим, что изменениям подвергается лишь один компонент К-знаний - лингвистические знания, причем не все, а лишь те, которые соответствуют наиболее подвижной части естественного языка - лексикону. Как правило, такие изменения заключаются в пополнении базы знаний, например, в создании новой словарной статьи для слова, впервые встретившегося в тексте и не знакомого системе.

Знания, отображающие требования семантической корректности и простоты интерпретации, общеязыковые и внешние нормы, может изменять только администратор системы.

Для внесения изменений в базу лингвистических знаний используются сервисные программы; для изменения текста - подсистема редактирования ЛИНАРа.

Отметим, что (даже при работе с ЛИНАР в диалоговом режиме) редактирование текста обычно производится по завершении работы программ контроля. Это связано с тем, что исправление фиксируемых системой ошибок и неточностей зачастую требует переделки относительно больших фрагментов текста (разбиение длинной фразы на несколько более простых, устранение неоднозначности трактовки и т.п.). Однако некоторые - локальные - изменения можно внести в текст сразу же в момент обнаружения ошибки. Поэтому в ряде программ контроля, например, в программах орфографического уровня, предусмотрена возможность исправления фиксируемых ошибок в момент их обнаружения.

Программы контроля

Программы контроля текста могут быть классифицированы по нескольким критериям.

Первый критерий связан с анализируемым программой аспектом текста. В соответствии с этим критерием выделяются следующие группы программ одноаспектного контроля:

- контроль орфографии (включая поиск ошибок в склонении и спряжении слов);
- анализ лексического состава текста;
- стилистический контроль;
- проверка выполнения правил структуризации текста;
- контроль синтаксической структуры;
- пунктуационный контроль;
- семантический контроль.

По второму критерию программы одноаспектного контроля подразделяются на программы локального и глобального анализа текста. Программы первой группы обрабатывают мелкие фрагменты текста: отдельные словоформы, словосочетания, специальные символы, не исследуя их контекстные связи или ограничиваясь учетом ближайшего окружения (соседнего слова справа, например). Локальный анализ характерен для программ орфографического, лексического и (частично) стилистического контроля. Программы, осуществляющие глобальный анализ, исследуют, как правило, структуру более крупных единиц текста: фраз и иногда абзацев (синтаксический и семантический контроль), текста в целом.

Третий критерий связан с характером результата, получаемого программой одноаспектного анализа. Основная часть программ контроля обнаруживает те или иные несоответствия текста и К-

знаний, используемых в текущем сеансе. Результатом их работы является список выявленных несоответствий (нарушений). Однако некоторые программы, как уже отмечалось, определяют отдельные свойства текста, не оценивая их. Так, программа ЛЕКС1 составляет частотный словарь исследуемого текста (фрагмента текста). Оценку полученным результатам дает человек - пользователь ЛИНАР, он же принимает решение о дальнейших действиях. Его реакция может быть, например, такой - поработать над текстом пункта 4.5.1., поскольку в этом тексте (занимающем всего две страницы) 26 раз встречается слово *знания* (в различных формах) и 7 раз - слово *соответственно*.

Только что рассмотренный пример (программа ЛЕКС1) можно использовать и для иллюстрации четвертого критерия классификации программ контроля. Эта программа, как и ряд других, выдает некоторую глобальную информацию об исследуемом фрагменте текста, не фиксируя, в каких позициях (абзацах, фразах или строках) были обнаружены в тексте формы различных слов. Другие программы, например программы проверки орфографии, локализуют обнаруживаемые ими свойства (дефекты) текста.

И наконец, отметим еще одно (формальное) различие программ контроля. Для всех программ основным параметром является подлежащий обработке фрагмент текста. Однако для некоторых программ нужно обязательно указать дополнительные параметры, конкретизирующие задание. Например, при вызове программы ЛЕКС2 нужно указать, какие именно грамматические признаки слов интересуют пользователя.

Некоторые программы контроля получают в качестве параметра предельно допустимые (пороговые) числовые значения количественно оцениваемых параметров текста. Отметим, что, меняя порог, можно варьировать уровень требований, предъявляемых к тексту, моделируя тем самым оценку его разными адресатами. Например, можно установить в качестве предельно допустимой длины фразы 25 слов или ограничить число придаточных предложений (в составе сложного предложения) двумя. Фразы, в которых эти пороговые значения превышены, будут классифицированы соответствующими программами контроля как недопустимые.

Орфографический контроль

Программы орфографического контроля обнаруживают (и предлагают варианты исправления) мотивированные грамматические ошибки в основах и окончаниях (флексиях) слов, записанных в словарь системы, и слов, встретившихся ей впервые (незнакомых), а также случайные, или немотивированные, ошибки.

Основные классы учитываемых случайных ошибок таковы:
- пропуск одной буквы (*а**с**емблер*),
- одна лишняя буква (*авт**т**окод*),
- замена одной буквы (*ко**н**пьютер*),
- перестановка двух соседних букв (*аг**л**оритм*).

Признаком ошибки служит появление в обрабатываемом тексте формы незнакомого слова.

Предпринимается попытка «свести» такое незнакомое слово к знакомому с помощью преобразований, обратных перечисленным выше (считается, что ошибка могла возникнуть в результате одного из таких «прямых» преобразований знакомого слова). Для предварительной оценки близости слов (основ слов) используется специально разработанная метрика.

Одна из программ обнаруживает ошибки в датах, задаваемых в тексте с помощью конструкций вида ДД.ММ.ГГ. Если задан и диапазон возможных дат, проверяется также принадлежность всех представленных в исследуемом тексте дат этому диапазону.

Примеры работы программ:

```
прочитанна - ОШИБКА В СЛОВОИЗМЕНЕНИИ !
ОЖИДАЕМОЕ СЛОВО: прочитана
рассчета   - ВОЗМОЖНА ОШИБКА ТИПА "удвоение буквы"
ОЖИДАЕМОЕ СЛОВО : расчета
10.25.89.
ОШИБКА В ДАТЕ          - недопустимая дата: месяц: 25
```

Анализ лексического состава текста

Программа ЛЕКС1

Программа подсчитывает, сколько раз в тексте (области) употребляется то или иное слово. Программа формирует полный список всех различных слов текста с указанием частот их

встречаемости. Можно задать диапазон частот (например, от 10 до 20 вхождений или ровно 15 вхождений) и сформировать список слов, количество употреблений которых лежит в границах этого диапазона. Если диапазон не задан, формируется полный частотный словарь текста.

Программа ЛЕКС2

Программа формирует список слов, обладающих указанными лексико-грамматическими характеристиками, например, находит все существительные, все причастия или все аббревиатуры, встретившиеся в тексте (области). Слова упорядочиваются по алфавиту, для каждого слова подсчитывается число его вхождений в исследуемый текст. Программа предназначена для анализа словарного состава текста.

Программа ЛЕКС3

Программа находит все вхождения в исследуемый текст (область) любых форм указанного (ключевого) слова и для каждого вхождения выдает контекст установленной длины - цепочку слов, находящихся от ключевого слова на расстоянии, не превышающем заданную длину. Программа удобна для анализа лексического состава текста и контроля используемых терминов и терминологических словосочетаний.

Программа ЛЕКС4

Программа находит в исследуемой области текста все слова, не входящие в формируемый в начале очередного сеанса словарь системы ЛИНАР, - т.е. слова, не знакомые очередному адресату. Для исправления текста следует либо заменить обнаруженные слова синонимами, либо расширить словарь системы. Возможно, что некоторые из обнаруженных слов являются известными системе словами, введенными с ошибками.

Программа ЛЕКС5

Программа осуществляет поиск каждой из обнаруживаемых в тексте (области) аббревиатур последовательно в трех списках:

N 3 – списке аббревиатур, вводимых непосредственно в тексте (этот список формируется динамически самой программой ЛЕКС5);

N 2 – списке, формируемом в начале работы с текстом на основе перечня используемых сокращений;

N 1 – словаре общепринятых сокращений.

В списке N 1 поиск ведется в последнюю очередь так как он, во-первых, самый большой, и во-вторых, если, например, в списках N 3 и N 1 присутствует одно и то же сокращение, но с различными расшифровками, то приоритет имеет сокращение из списка N 3. Результатом работы является список используемых в тексте аббревиатур с указанием их локализации в тексте и типа аббревиатуры.

Программа ЛЕКС6

Программа осуществляет контроль за переопределением известных системе аббревиатур. Если, например, в разделе 1.2. встретилась аббревиатура СВП (с расшифровкой в тексте – *«схема внешних прерываний»*), а в списке N 2 аббревиатура СВП сопоставлена термину *«субкомплекс внешней памяти»*, фиксируется ошибка: недопустимое переопределение аббревиатуры из перечня.

Программа ЛЕКС7

Программа проверяет правильность расшифровки, то есть тот факт, что аббревиатура читается в расшифровке по началам слов, причем некоторые слова расшифровки могут не участвовать в образовании аббревиатуры.

Пример работы программы:

```
Эта организация - центр переводов (ВЦП).
НЕСООТВЕТСТВИЕ АББРЕВИАТУРЫ И РАСШИФРОВКИ:
ВЦП - центр переводов
```

Программа ЛЕКС8

Программа ЛЕКС8 (без параметров) проверяет правильность оформления списка используемых в тексте аббревиатур (для отчета по НИР - это «Перечень условных обозначений,

символов, единиц и терминов»). Предполагается, что каждая пара ***аббревиатура - расшифровка*** в перечне представлена одной строкой. В процессе обработки перечня заполняется список замечаний.

Пример работы программы:

```
ОБРАБАТЫВАЕТСЯ ПЕРЕЧЕНЬ АББРЕВИАТУР:
 БНК - бортовой нейрокомпьютер
 БНФ - бекусовская нормальная форма
 КПД - канал прямого доступа
 ОЗУ
 МПК - микропрограммируемый контроллер
 ОРЗ - общий регистр записи
 ПНП - перейти в неустойчивое положение
 СВП - субкомплекс внешней памяти
 СПТ - субкомплекс рабочего таймера
ЗАМЕЧАНИЯ:
4 : ОЗУ  * НЕТ РАСШИФРОВКИ
5 : МПК  * НАРУШЕНИЕ АЛФ. ПОРЯДКА
7 : ПНП  * РАСШИФРОВКА НЕ ЯВЛЯЕТСЯ ГРУППОЙ СУЩЕСТВИТЕЛЬНОГО
9 : СПТ  * НЕСООТВ: АББР.-РАСШ.
```

Стилистический контроль

Программы данного блока фиксируют внешние характеристики фраз, свидетельствующие о сложности их структуры, а следовательно, и о сложности восприятия смысла. Имеются, например, программы, контролирующие длину фраз, количество запятых, количество придаточных предложений, наличие во фразах текста длинных цепочек слов в родительном падеже (например, *значений аргументов программы пользователя*) или цепочек однокоренных слов (*пользователь может воспользоваться, транслятор транслирует*). Есть программы контроля стилистической окраски слов. В научно-технической литературе нежелательно употребление устаревших слов и канцеляризмов (*ибо; вышепоменованный*), жаргонизмов (*виндуза*), разговорных оборотов (*этот алгоритм, уж поверьте; . . .*). При обнаружении таких слов в тексте их рекомендуется убрать или заменить более нейтральными синонимами. Особый класс составляют слова, явно характеризующие специфику темы (предметной области), раскрывать которую иногда нежелательно. Например, в документе для внутреннего пользования можно употребить термин *военно-космический*, а в тексте сообщения, передаваемого по открытым каналам связи его целесообразно заменить (соответствующая программа предлагает слово-замену *специальный*).

Контроль структуры текста

Данные программы контролируют правильность оформления отдельных структурных частей текстового документа с точки зрения соответствующих нормативных требований (например, требований ГОСТа 7.32-81, регламентирующего правила оформления научно-технического отчета). Проверяется оформление титульного листа, списка исполнителей, реферата и других разделов документа.

Синтаксический контроль

Программа СИНТ1

Программа СИНТ1 находит в указанной области именные словосочетания вида <прилагательное> + <существительное> и <существительное> + <существительное в форме родит. падежа> и др. Программа может оказаться полезной при анализе лексического состава текста и при поиске терминологических словосочетаний, особенно в тех случаях, когда различные фрагменты текста написаны разными авторами (возможно, использующими близкие, но не совпадающие термины). Найденные программой словосочетания группируются вокруг «ключевого слова» - существительного, играющего роль синтаксической вершины словосочетания.

Ряд программ синтаксического контроля обнаруживает нарушения обычного (нейтрального) порядка слов и взаимного расположения групп слов. Такие нарушения могут затруднить восприятие текста.

Например: «*Раздел второй посвящен описанию новых алгоритмов».* или *«Использует этот алгоритм всего две вспомогательные переменные»*.
Отметим, что иногда нарушение нейтрального порядка слов может намеренно использоваться автором текста с целью изменения логического ударения, усиления (*«Алгоритм этот очень эффективен!»*).

Программа СИНТ2

Программа СИНТ2 осуществляет контроль придаточных предложений с союзным словом *который*, а именно, проверяет однозначность установления связи между союзным словом и его словом-хозяином из главного предложения. В случае, когда таких слов-хозяев не обнаружено или их более одного, выдается соответствующая диагностика.

Пример работы программы:

```
Рассмотрим структуру памяти вычислительной машины, в которой хранятся
команды.
СЛОВО которой ИМЕЕТ БОЛЕЕ ОДНОГО СЛОВА-ХОЗЯИНА В
ГЛАВНОМ ПРЕДЛОЖЕНИИ: машины, памяти, структуру
Каждому каналу соответствует свое устройство, которые в свою очередь
связаны с главной ЭВМ.
СЛОВО которые НЕ ИМЕЕТ СЛОВА-ХОЗЯИНА В ГЛАВНОМ ПРЕДЛОЖЕНИИ
Мощь языка Си - результат выявления его авторами потребностей
программистов, которые возникают при программировании на языке
ассемблера.
СЛОВО которые ИМЕЕТ БОЛЕЕ ОДНОГО СЛОВА-ХОЗЯИНА В ГЛАВНОМ ПРЕДЛОЖЕНИИ:
программистов, потребностей, авторами
```

Пунктуационный контроль

Пунктуационные ошибки в реальных предложениях русского языка встречаются довольно часто.

Блок пунктуационного контроля системы ЛИНАР разработан на основе весьма полной пунктуационной модели русского языка. Полнота и корректность базовых знаний является основой достижения устойчивости и эффективности программных средств, реализованных на основе данной модели. В то же время блок пунктуационного контроля является «открытым», т.е. построен таким образом, чтобы обеспечить возможность работы средств адаптации и, при необходимости, введения новых правил пунктуации.

При проверке пунктуации можно использовать любое количество программ контроля, выбирая их при этом по различным признакам. Например, можно осуществлять проверку только тех правил, которые выявляют лишние знаки препинания, можно только тех, которые выявляют пропущенные знаки препинания и т.д. При подобной настройке может меняться совокупность пунктуационных правил, степень жесткости требований по соблюдению каких-либо условий и т. д., что позволяет оценивать качество текста с точки зрения различных категорий пользователей. Набор желаемых для данного сеанса модулей формируется в начале работы пользователем.

Пример работы программ пунктуационного контроля:

```
В ПРЕДЛОЖЕНИИ:
Только и развлечений , что кино раз в неделю
ЗАМЕЧЕНА ПУНКТУАЦИОННАЯ ОШИБКА.
В выделенном месте не должно быть данного знака препинания. В
рассматриваемом случае запятая перед что не ставится .
Необходимо пояснение ошибки? (Д/Н)
     Д
В безглагольном предложении перед союзом что в выражении только и …
что , за которым следует имя существительное или местоимение, запятая
не ставится.
Необходимы примеры правильного применения данного правила? (Д/Н)
     Д
Только и денег что пятак в кармане.
Только и разговоров что о них двоих.
```

Семантический контроль

Программа СЕМ1

Программа обнаруживает несовпадение ожидаемых семантических признаков актантов (подлежащее, дополнения) глагола и признаков слов (групп слов), реально занимающих соответствующие позиции. Такое несовпадение мешает завершить анализ фразы, поскольку синтаксически допустимая связь не может быть установлена из-за семантических противоречий. Проверяя употребление в тексте глаголов, программа обращает внимание пользователя на «подозрительные» актантные конструкции.

Пример работы программы:

```
Все рассматриваемые программы написаны на ассемблере.
НЕСОВПАДЕНИЕ СЕМАНТИЧЕСКИХ КЛАССОВ!
В ОПИСАНИИ ГЛАГОЛА "написать" СЕМ.-КЛАСС АКТАНТА:
=язык_программирования=
РЕАЛЬНЫЙ АКТАНТ ассемблере ИМЕЕТ СЕМ.-КЛАСС: =транслятор=
Схема прерываний подключается к магистрали.
НЕСОВПАДЕНИЕ СЕМАНТИЧЕСКИХ КЛАССОВ!
В ОПИСАНИИ ГЛАГОЛА "подключаться" СЕМ.-КЛАСС АКТАНТА:
=устройство=
РЕАЛЬНЫЙ АКТАНТ схема прерываний ИМЕЕТ СЕМ.-КЛАСС:
=структура2=
```

Программа СЕМ2

Программа проводит полный синтактико-семантический анализ фраз указанной области текста. При этом фиксируются случаи, когда фраза имеет (в контексте предметной области, к которой относится текст) более одной интерпретации, т.е. допускает неоднозначное толкование.

Пример работы программы:

```
Снижение напряжения вызвало отключение принтера.
НЕОДНОЗНАЧНАЯ ИНТЕРПРЕТАЦИЯ!
  1 трактовка:
=причина=     : снижение напряжения
=следствие=   : отключение принтера
  2 трактовка:
=причина=     : отключение принтера
=следствие=   : снижение напряжения
```

Программа СЕМ3

Программа СЕМ3 проверяет однозначность установления связи между личным местоимением и словом, на которое ссылается это местоимение. Если такое слово не найдено или же таких слов более одного, выдается соответствующая диагностика.

Пример работы программы:

```
Каждому каналу сопоставлено определенное устройство. Они, в свою
очередь, связаны с главной ЭВМ.
ДЛЯ МЕСТОИМЕНИЯ они В ПРЕДШЕСТВУЮЩЕЙ ФРАЗЕ НЕ НАЙДЕНО СЛОВ,
НА КОТОРЫЕ ЭТО МЕСТОИМЕНИЕ ССЫЛАЕТСЯ
Рассмотрим структуру памяти ЭВМ. Она состоит из двух основных частей.
ДЛЯ МЕСТОИМЕНИЯ она В ПРЕДШЕСТВУЮЩЕЙ ФРАЗЕ НАЙДЕНО БОЛЕЕ ОДНОГО СЛОВА,
НА КОТОРОЕ ССЫЛАЕТСЯ ЭТО МЕСТОИМЕНИЕ:  ЭВМ, памяти, структуру
```

Программа СЕМ4

Программа проверяет, принадлежат ли значения количественно оцениваемых свойств описываемых в тексте объектов заданному диапазону. В случае, если значение свойства выходит за границы диапазона, процедура выдает соответствующую диагностику.

Пример работы программы:

```
Информация передается в сопроцессор АК-34 по 16 каналу.
ОБ'ЕКТ: сопроцессор АК-34        ГРУППА: 16 каналу
ВЫХОД ЗНАЧЕНИЯ ЗА ВЕРХНЮЮ ГРАНИЦУ ДИАПАЗОНА
(СОПРОЦЕССОР АК-34 ИМЕЕТ КАНАЛЫ: 0,1,2, ... 15)
```