

РАЗРАБОТКА И ПРОГРАММНАЯ РЕАЛИЗАЦИЯ ПРИЛОЖЕНИЯ ДЛЯ УНИКАЛИЗАЦИИ ТЕКСТОВОЙ ИНФОРМАЦИИ

Белый А.М., Головина Д.А., Милованов М.М.

ФГБОУ ВПО «Сибирский государственный индустриальный университет»,
г. Новокузнецк, Россия

Использование современных программных средств, предназначенных для работы с текстами, является очень актуальным вопросом для всей индустрии печати. Использование уникальных текстов дает неоспоримые преимущества как автору текста, так и читателю. Для решения проблемы уникальности текстов проведен анализ существующих средств и разработано приложение, позволяющее создавать уникальные тексты для публикации. Произведен анализ уникальности, а также проведено тестирование. Полученные результаты позволяют сделать вывод о возможности применения разработанных алгоритмов для унификации текстов статей и иных публикаций.

Ключевые слова: синонимайзер, анализ данных, уникальность текста, синонимы, части речи, существительное, прилагательное, глагол, СУБД.

The using of modern software tools designed to work with texts, is a very important issue for the entire printing industry. Using the unique texts give great benefits, as the author of the text and the reader. An analysis of existing tools was conducted to solve the problem of uniqueness of texts, also application for creating unique texts for publication was developed. The analysis of uniqueness made, and testing made. The results obtained suggest the possibility of using the developed algorithms for unification text articles and other publications.

Keywords: synonymizer, data analysis, the uniqueness of the text, synonyms, parts of speech, noun, adjective, verb, DBMS.

Как правило, программы для уникализации текста имеют базу данных для хранения синонимов, при помощи которой осуществляется замена слов. Также эти программы могут работать в автоматическом и ручном режимах, тем самым предоставляя пользователю возможность выбора синонимов из списка. При этом предусмотрена возможность занесения пользователем как новых слов, так и пар «слово-синоним» в базу данных.

Текст для обработки может быть введен в систему при помощи клавиатуры или же загружен из файла. После преобразования полученный текст может быть сохранен в текстовый файл.

Задачи, решаемые разрабатываемым программным продуктом:

1. Создание базы данных слов с возможностью хранения их атрибутов таких как род, число, падеж и т.д.
2. Реализация механизма замены слов.
3. Реализация механизма преобразования текста в соответствии с правилами русского языка.
4. Реализация механизма сопоставления исходного и полученного текстов.
5. Реализация механизма сохранения и загрузки текстов.

Источником данных для системы является СУБД, созданная в среде InterBase версии 7.5.

Основой для алгоритма работы программы является грамматический разбор слов в соответствии с правилами русского языка. Приложение «умеет» подбирать синонимы только к словам трех частей речи: существительное, прилагательное и глагол. После отнесения конкретного слова к какой-либо части речи к нему подбирается синоним, из уже имеющихся в базе данных, относящийся к этой части речи. Алгоритм уникализации текста является авторским и представлен на рис. 1.

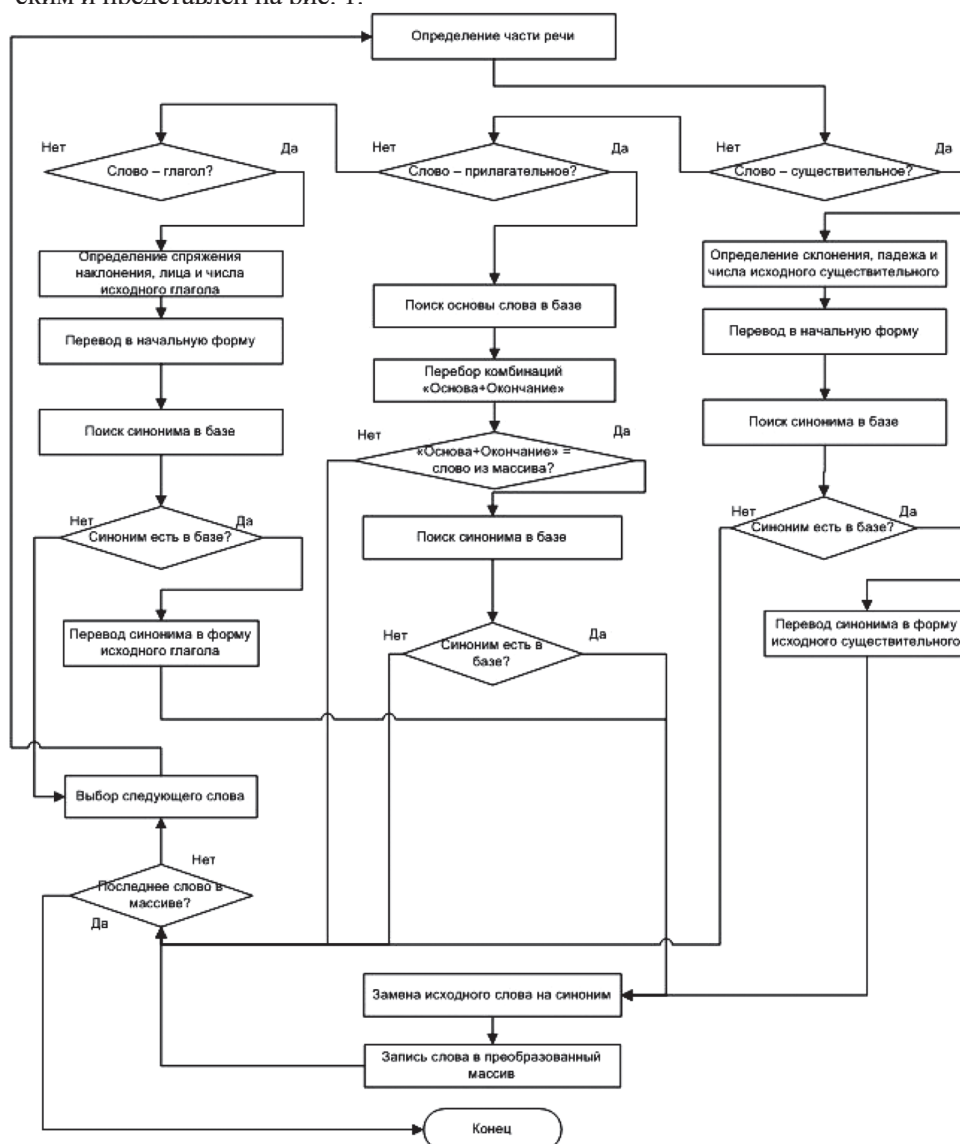


Рис. 1. Алгоритм работы программы

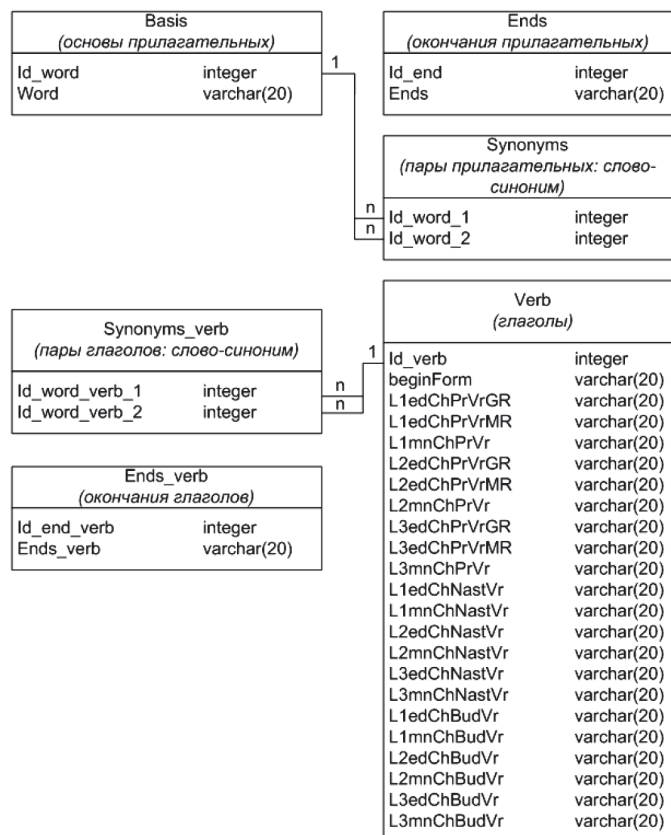


Рис. 2. Структура базы данных

Основным языком программирования для реализации данного проекта выбран язык Pascal и SQL, поэтому необходимо использовать официальную документацию к данным языкам [3]. На рис. 2 приведена структура базы данных, предназначенной для хранения информации о словах и синонимах. В программе используются классы для представления слов в виде одной из следующих частей речи: существительное, прилагательное, глагол. Диаграмма классов представлена на рис. 3.

Предполагается, что пользователи смогут работать с приложением только в однопользовательском режиме, т.е. у каждого пользователя будет персональная информационная база с уникальными синонимами. Это позволит использовать приложение без дополнительных затрат. Также приложение обладает интуитивно понятным интерфейсом.

Для тестирования программы был взят отрывок из романа Л.Н. Толстого «Война и мир»:

«Толпа подошла к большому столу, у которого, в мундирах, в лентах, седые, плешивые, сидели семидесятилетние вельможи-старики, которых почти всех, по домам с шутами и в клубах за бостоном, видал Пьер. Толпа подошла к столу, не переставая гудеть».

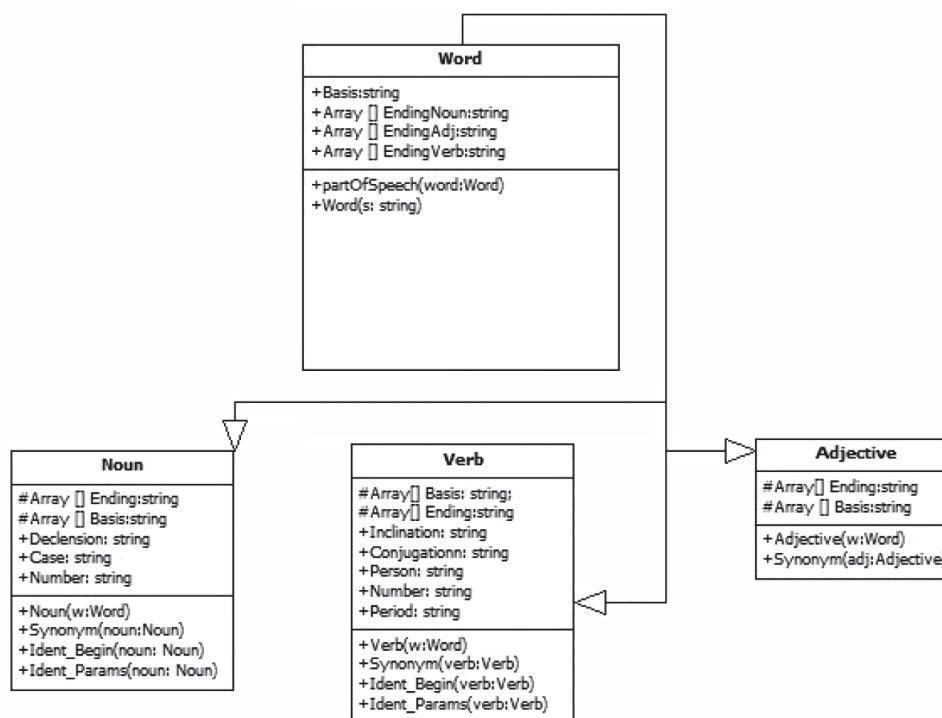


Рис. 3. Диаграмма классов

При тестировании текст как загружался из файла, так и вводился с клавиатуры в поле ввода. На этапе ввода текста не было выявлено отклонений в работе программы.

При нажатии на кнопку «Преобразовать» был получен следующий текст: «толпа приблизилась к огромному столу, у которого, в мундирах, в лентах, седые, лысые, расположились семидесятилетние вельможи-старики, которых почти всех, по домам с шутами и в клубах за бостоном, наблюдал пьер. толпа приблизилась к столу, не переставая шуметь».

В результате работы программы была произведена замена слов исходного текста на синонимы, имеющиеся в базе. Результат представлен на рис. 4.

При анализе результатов работы программы были выявлены следующие пары исходных слов и синонимов: подошла – приблизилась; большому – огромному; плешивые – лысые; сидели – расположились; видал – наблюдал; гудеть – шуметь.

Все слова заменились на синонимы с сохранением исходной формы, согласованность слов в полученном тексте не нарушена.

При нажатии на кнопку «Ручной режим» открывается диалоговое окно, позволяющее выбрать в выпадающем списке слово из исходного текста и выбирать для него любой синоним из имеющихся в базе.

Вместо синонима «огромному» к слову «большому» был подобран синоним «крупному», в результате чего преобразованный текст изменился.

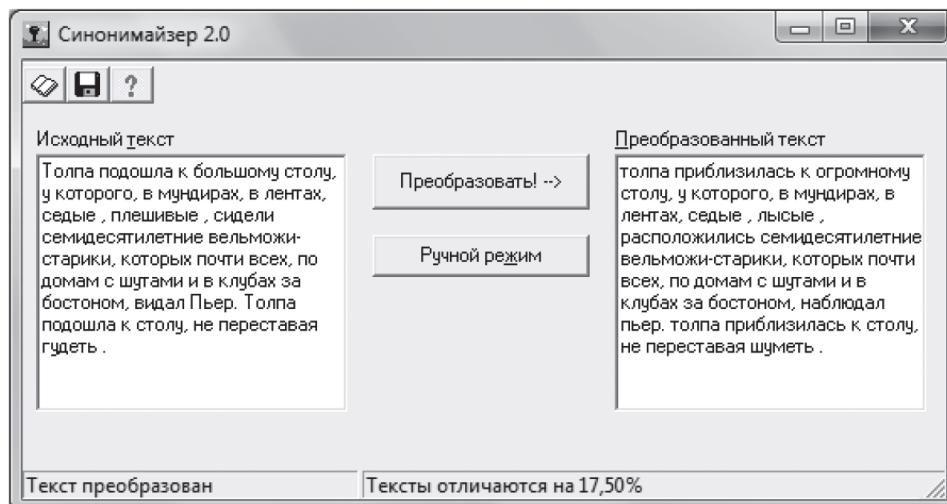


Рис. 4. Результат преобразования текста

Исходный и преобразованный текст отличаются на 17,50 %. Общее количество слов в тексте без учета пробелов, с учетом знаков препинания, отделенных пробелами: 40. Количество слов, замененных на синонимы: 7. Процент отличия текстов был вычислен самостоятельно, для проверки правильности подсчетов программы:

$$(7/40)100 \% = 17,50 \%$$

На этапе сохранения преобразованного текста в файл не было выявлено отклонений в программе.

Список использованных источников

1. ГОСТ 19.201-78 «Техническое задание. Требования к содержанию и оформлению».
2. ГОСТ 34.602-89 «Техническое задание на создание автоматизированной системы»
3. URL: <http://www.ibase.ru/ibfaq.htm> (дата обращения: 26.12.2014).
4. Тидвелл Д.: Разработка пользовательских интерфейсов. «Питер», 2008, 416 с.
5. Милованов М.М. Современные подходы к моделированию и анализу бизнес-процессов предприятия [Электронный ресурс] // «Управление экономическими системами. Электронный научный журнал», 2011, № 11. – Режим доступа: <http://www.uecs.ru/instrumentalnii-metody-ekonomiki/item/821-2011-11-30-11-53-58> (доступ свободный) – Загл. с экрана. – Яз. Рус.