OSTIS-2011
(Open Semantic Technologies for Intelligent Systems)

УДК [004.522+004.934]:004.89

СИСТЕМА РЕДАКТИРОВАНИЯ И ПОПОЛНЕНИЯ СЛОВАРЕЙ РЕЧЕВОГО ИНТЕРФЕЙСА ВОПРОСНО-ОТВЕТНОЙ СИСТЕМЫ ДЛЯ БЕЛАРУСКОГО И РУССКОГО ЯЗЫКОВ

Ю. С. Гецевич (*mix1122@gmail.com*)
В. Н. Вяльцев (*vltsvn@gmail.com*)
Объединенный институт проблем информатики НАН Беларуси, г.Минск, Республика Беларусь

Аннотация: Данная статья рассказывает о возможностях добавления новых и редактировании ранее введенных слов и их лексико-грамматических категорий в электронные словари для беларуского и русского языков, чтобы синтез речи по тексту мог корректно озвучить для пользователя любой ответ, сгенерированный семантической вопросно-ответной системой.

Ключевые слова: лексико-грамматические категории, программа добавления и редакции слов, семантическая вопросно-ответная система, система синтеза речи по тексту, электронный словарь.

Введение

Система синтеза речи Multiphone [Лобанов и др., 2008] является составной частью семантической вопросно-ответной системы [Житко и др., 2010]. Искусственно-речевой интерфейс принимает текстовые данные на вход и старается ответить пользователю в виде понятного текстового послания, которое озвучивается через блоки синтеза речи. Одним из самых важных блоков синтеза речи является лингвистический процессор. Он использует грамматический словарь для расстановки ударений и лексико-грамматических категорий в словах.

Всего в многоязычной системе синтеза речи [Lobanov et al., 2006] электронные грамматические словари содержат около 2,5 миллионов записей для беларуского языка (построен на основе словаря М. В. Бирило [Бірыла, 1987]), и около 2,5 миллионов записей для русского языка (построен на основе словаря А. А. Зализняка [Зализняк, 1980]). Эти электронные словари на высоком уровне позволяют решить проблему определения грамматических характеристик слов. Но для того, чтобы словари пополнялись новыми словами и, чтобы в словарях слова содержали правильную информацию о словах, необходима система редактирования и пополнения грамматических словарей.

1. Описание системы редактирования и пополнения грамматических словарей

Грамматический словарь построен в виде таблицы, содержащей записи слов для беларуского языка (табл. 1а) и для русского языка (табл. 1б) с обозначенными для них через тэги лексико-грамматическими категориями (ЛГК), ударениями (обозначается знаком «+» или «=») и приоритетами получения слов из словаря лингвистическим процессором.

Таблица 1а,б – Фрагменты электронного грамматического словаря с обозначением ударений (+), тегов и приоритетов слов для существительных.

Слово_Тег _Приоритет
…
зака+зчык_NNAMO_1
зака+зчыка_NNAMG_1
зака+зчыку_NNAMD_1
зака+зчыка_NNAMA_1
зака+зчыкам_NNAMI_1
…

(а) – беларуский словарь

Слово_Тег _Приоритет
…
зака+зчик_NCAMSN_1
зака+зчика_NCAMSG_1
зака+зчику_NCAMSD_1
зака+зчика_NCAMSA_1
зака+зчикам_NCAMPD_1
…

(б) – русский словарь

Для словаря разработан специальный программный интерфейс – AddWords (AB), как для чтения (через класс CVocReader), так и для пополнения, удаления и редактирования (через класс CVocEditor) его данных (Рис. 1).

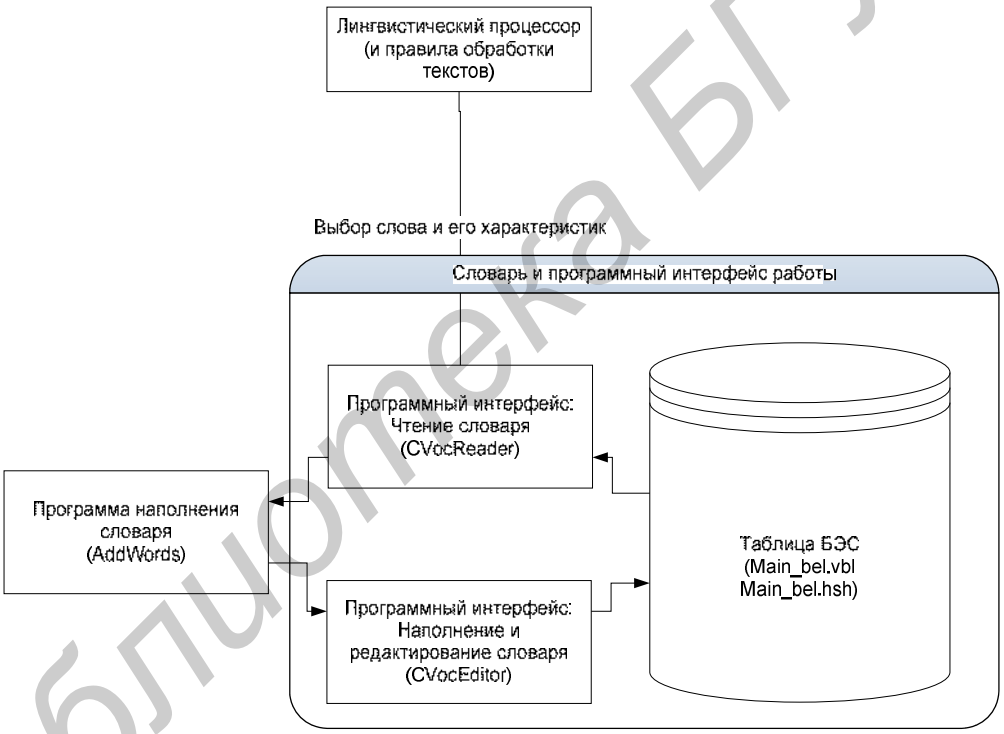


Рисунок 1 – Взаимодействие грамматического словаря с лингвистическим процессором и программой пополнения словаря

Если необходимо, чтобы некоторый неизвестный текст хорошо звучал на выходе всего синтеза речи, то нужно, чтобы все слова имели проставленные ударения, а также чтобы была снята омография слов в тексте. Программа АВ задумана, во-первых, для работы с текстами, которые обрабатываются лингвистическими алгоритмами с обращением к грамматическому словарю, а, во-вторых, для работы со словарем напрямую, чтобы узнавать, какие и как записаны слова в словаре. Мы будем рассказывать про программу АВ в обозначенном выше порядке целей ее разработки.

2. Работа с текстами в программе АВ

Перед обработкой текста программой АВ нужно установить необходимые параметры для лингвистического процессора (ЛП), используемого программой для обработки текстов. Через

команду «Open Lang\» можно выбрать необходимую папку с языковыми ресурсами, а через список выбора языков – конкретные языковые ресурсы: беларуские (Belarusian) или русские (Russian) (Рис. 2).

Важно отметить, что программа АВ может обрабатывать тексты даже с неправильно выбранными ресурсами, т.к. алгоритмы лингвистических обработок почти одни и те же для разных языков. Вследствие этого в данной работе примеры иногда будут приводиться только для одного языка.

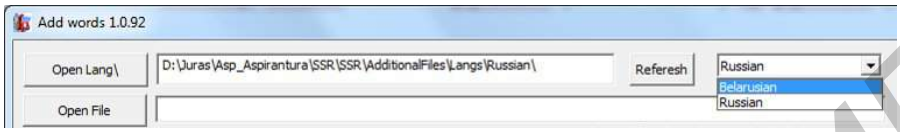
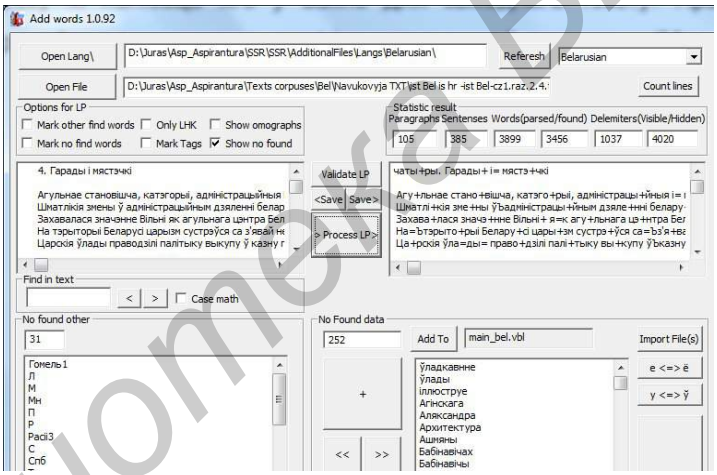


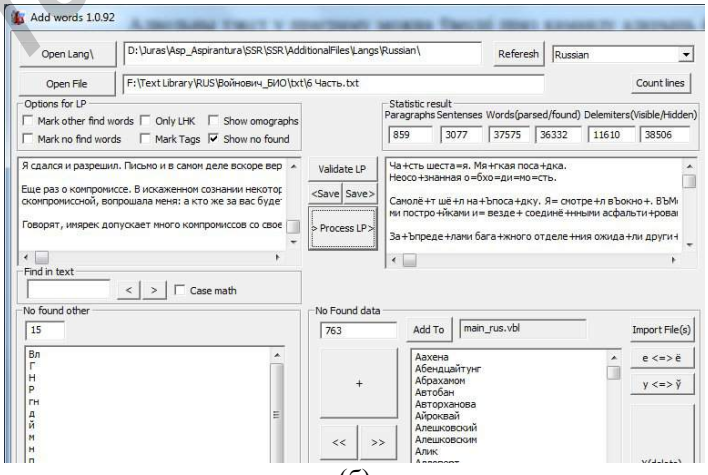
Рисунок 2 – Выбор языковых ресурсов для лингвистического процессора

Произвольный текст в программу можно ввести через команду «Открыть файл» (“Open File”) или сразу же написать его в окне для исходного текста (Рис. 3а, б). Через команду “Process LP” ЛП обрабатывает данные, результаты помещаются в три окна:

- Обработанный текст (правое верхнее окно)
- Нераспознанные выражения (левое нижнее окно) (No found other)
- Распознанные ненайденные выражения (правое нижнее окно) (No found data)



(а)

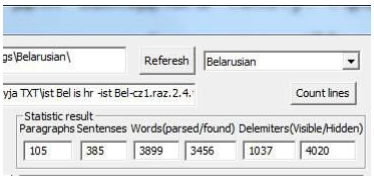


(б)

Рисунок 3 – Примеры обработки беларуского (а) и русского (б) текста и распределения результатов обработки с возможностью сохранения в файлы через команды “Save”

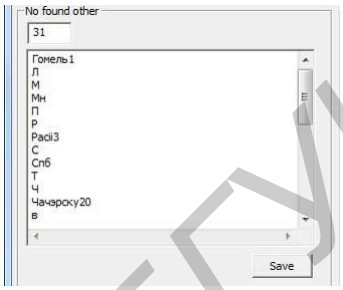
Полученные результаты обработки можно сохранять в текстовые файлы через команды кнопок “Save”, которые есть возле каждого текстового окна. Имена файлов для результатов обработки программа предлагает сохранять с соответствующими постфиксами: _result, _errorOther, _errorWords.

После обработки текста можно посмотреть разнообразную статистику по его структуре (Рис. 4): программа выводит в сгруппированной области “Statistics Result” количество параграфов, предложений, слов (обработанных и найденных), разделителей (видимых и невидимых) (Рис. 4а), в области “No found other” – количество нераспознанных выражений (Рис. 4б), в области “No Found data” – количество распознанных ненайденных выражений (Рис. 4в).



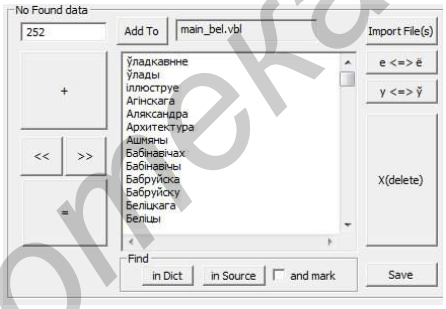
(а)

Количества параграфов, предложений, слов, разделителей



(б)

Количество нераспознанных выражений



(в)

Количество распознанных ненайденных выражений

Рисунок 4 – Вывод статистики по структуре текста

Пользователь может задавать специальные параметры для лингвистического процессора перед обработкой текста через выбор конкретных опций, которые находятся в области настроек для ЛП “Options for LP” (Рис. 5.).

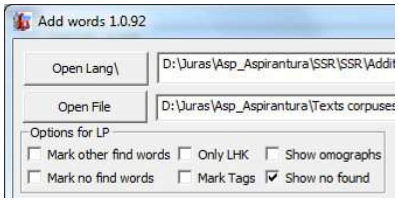


Рисунок 5 – Настройки обработки текста

Приведем описание возможных опций настроек обработки ЛП входного текста (Табл.2).

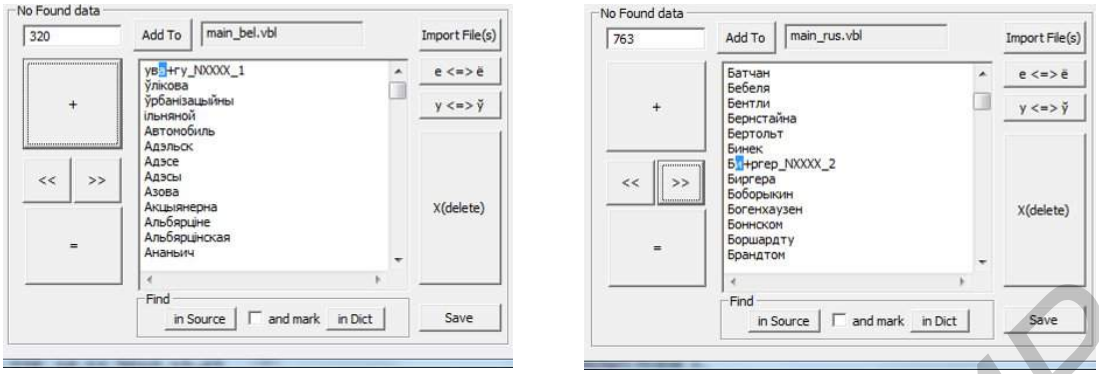
Таблица 2 – Описание настроек для ЛП в программе АВ

Опция настройки	Описание
Обозначить другие найденные слова (Mark other find words)	В окне обработанного текста выводит все найденные в словаре словоформы слова и омографы слова. *Например, слово “замок” преобразуется в замо+к{3, замо+к_NNIMA_1, замо+к_VPIPM_1, замо+к_NNIMO_1}.*
Обозначить ненайденные слова (Mark no find words)	В окне обработанного текста ненайденные слова в словаре обрамлять звездочками. *Например, слово с опечаткой “малоко” преобразуется в *ма=ло=ко=*.*
Показывать только ЛГК (Only LHK)	В окне обработанного текста показывать только названия частей речи слов. *Например, выражение “мама мыла дом” обрабатывается в “существительное глагол существительное”.*
Обозначать тэги (Mark Tags)	В окне обработанного текста показывать слова с отметкой частей речи через тэги и приоритета слова в словаре. *Например, выражение “беларуская шляхта” обрабатывается в выражение “белару+ская_JJFO_1 шля+хта_NNIFO_1”.*
Показывать омографы (Show omographs)	В окне ненайденных данных отражать слова омографы. *Например, в беларуском словаре у слова “казачку” существует 4 записи: ка+зачку_NNIFA_1, казачку+_NNAMR_1, казачку+_NNAMD_1, каза+чку_NNAFA_1, поэтому ЛП будет обрабатывать это слово по принципу омографов: если не стоит приоритет на каком-то конкретном слове, то в разных местах полных ударений будут поставлены частичные (“ка=за=чку=”) и слово будет выведено в списке распознанных, но ненайденных данных.*
Показывать ненайденные слова (Show no found)	В окне распознаных ненайденных данных будут отражаться ненайденные в словаре слова. *Например, слова “Воронков” нет в словаре, как и многих других имен собственных – значит это слово будет преобразовано в слово с частичными ударениями “Во=ро=нко=в” и будет выведено в окно ненайденных слов.*

По умолчанию ЛП настроен на обработку текста и вывод распознанных ненайденных слов, т.к. выбрана опция “Show no found”. Опишем возможности обработки ненайденных слов через программу АВ. Позже опишем возможность обработки найденных омографов в тексте, когда выбрана опция “Show omographs”.

2.1. Обработка распознанных ненайденных выражений

Для добавления новых слов с ударениями и ЛГК разработан специальный набор клавиш (Рис. 6а, б).



(а) (б)

Рисунок 6 – Интерфейс в программе АВ для обозначения ударений и ЛГК в новых словах

Клавишами со знаками стрелочек “<<” и “>>” можно перемещаться синим маркером по гласным незнакомых слов (и по букве ‘ў’ для беларуского языка) и ставить главное ударение ‘+’ или частичное ударение ‘=’ в словах нажатием соответствующих клавиш “+” і “=”, расположенных сверху и снизу клавиш со стрелками. Удаление ненужного ударения на гласной осуществляется через ‘X(delete)’.

Новый тег для слова обозначается выражением – ‘(слово_с_ударением)_тег’. Например, выражение “ува+гу_NXXXX_1” добавит в словарь слово “увагу” с полным ударением на вторую букву с тэгом NXXXX (существительное) и с приоритетом 1.

В некоторых словах с ‘е’ и ‘ў’ выделенные гласные ‘е’ или ‘ў’ могут быть быстро изменены пользователем на ‘ё’ или ‘у’ (или, соответственно, наоборот) клавишами ‘е⇔ё’, ‘у<=>ў’, что соответственно актуально для русских и беларуских текстов.

Клавиша “in Source” в сгруппированной области “Find” подсвечивает синим маркером конкретное слово во входящем тесте, чтобы пользователь мог корректно проставить характеристики слова в соответствии с контекстом слова в тексте (Рис. 7). Опция “and Mark” включает автоматическое слежение за неизвестными словами в тексте окна ввода.

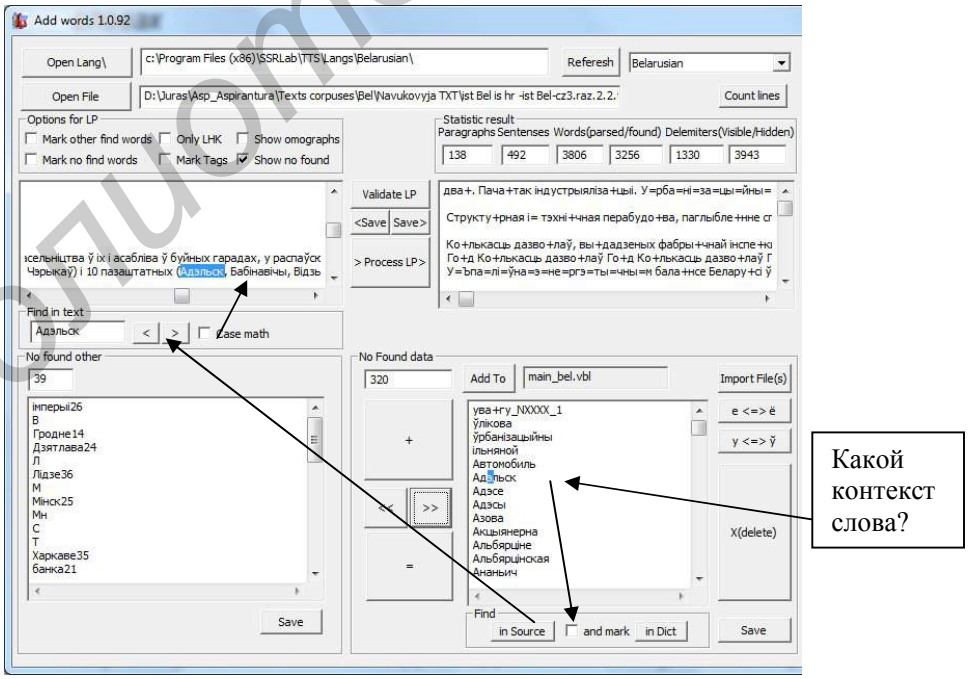


Рисунок 7 – Возможность увидеть контекст обрабатываемого неизвестного слова

Когда слова обозначены, их можно сохранить в промежуточный файл кнопкой “Save”, а обозначенные слова клавишей ‘Add to’ добавляются в конкретный словарь, например, в

“main_bel.vbl”. Когда нужно добавить большое количество слов через один файл или много файлов с новыми словами, в которых обозначены ударения и ЛГК, то можно воспользоваться клавишей ‘Import File(s)’. Будет вызван диалог предложения выбора файла или файлов словами (Рис. 8).

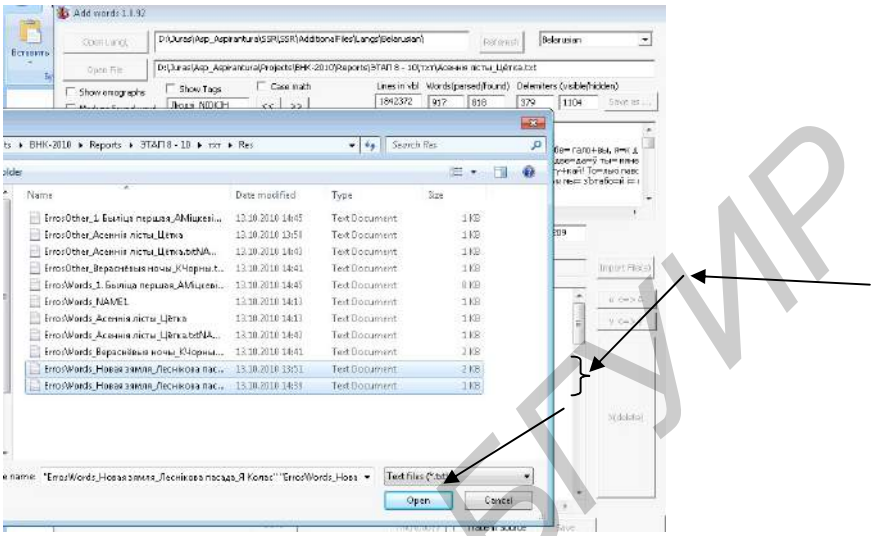


Рисунок 8 – Диалог выбора файла(ов) с новыми словами для словаря

Программа АВ предлагает добавить в словарь только слова с обозначенными ударениями на гласных буквах, причем, главное ударение должно быть только одно, частичных может быть несколько (Рис. 9). После согласия в словарь добавляются обозначенные слова и словарь готов для дальнейшего использования. Если еще раз обработать входной тест, то обозначенные и добавленные в словарь слова должны перестать показываться в окне распознанных ненайденных слов.

Рисунок 9 – Диалог предложения добавить новые обозначенные слова в словарь

2.2. Обработка нераспознанных выражений

Иногда в тексте имеются нераспознанные выражения для ЛП. Они возникают, во-первых, из-за того, что ЛП до конца еще не разработан для всех символьных случаев (например, для всех форматов даты, т.к. пользователь может использовать довольно много из имеющихся для разных стран), а, во-вторых, нет никаких ограничений на произвольный набор символов в подаваемом тексте, а догадаться об этой последовательности невозможно для общего случая. Поэтому для нераспознанных выражений предусмотрена только операция сохранения “Save” в файл (Рис. 4б) и дальнейшая передача разработчикам программы АВ для уточнения алгоритмов работы АВ.

2.3. Обработка омографов

Для поиска омографов в исходном тексте пользователь может выбрать опцию “Show omographs” в настройках ЛП и обработать входной текст нажатием на “Process LP”. Омографы появятся в окне “No Found data” (Рис. 11, а-путь «найти омографы»).

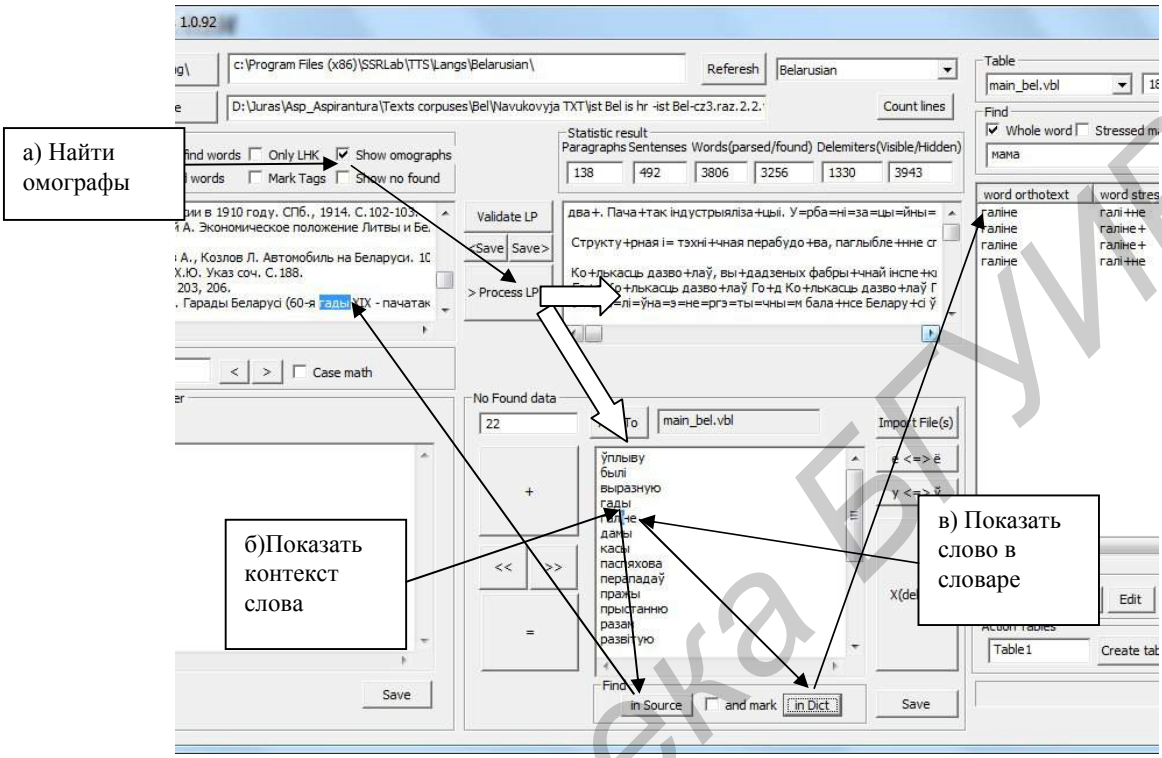


Рисунок 10 – Поиск и обработка омографов

Когда пользователь выберет конкретное слово, то ему важно знать контекст слова-омографа в тексте, который можно увидеть через описанную выше возможность клавиши “in Source” в сгруппированной области “Find (Рис. 11, б-путь «показать контекст слова»), и какие словоформы существуют в словаре (Рис. 11, в-путь «показать слово в словаре»).

Вследствие того, что программа считает словами-омографами слова, у которых одинаковый приоритет и разные главные ударения, то, изменив приоритет словоформы слова-омографа на более высокий, посредством работы напрямую со словарем, можно частично снять омографию слова в тексте.

3. Работа напрямую со словарем в программе AB

Активный грамматический словарь всегда имеет расширение *.vbl, и его можно увидеть в сгруппированной области “Table”. Рядом с ним отражается количество записанных в словарь слов (Рис. 12а,б).

Сгруппированная область “Find” позволяет выбрать необходимые настройки поиска слова в словаре (Табл. 3). Слово может искаться, как есть, так и с помощью масок поиска. Так можно найти слова с какой-либо вариативностью произвольных или конкретных гласных и согласных, ударных или безударных букв.

Таблица 3 – Описание настроек поиска слова в словаре (0 – означает, что опция не выбрана, да – выбрана)

Опция включена			Описание
Целое слово	С учетом ударения	Отличать прописные/строчные буквы	
да	0	0	Конкретное слово ищется в словаре с переводом прописных букв в строчные.
да	0	да	Конкретное слово ищется в словаре без перевода букв из прописных в строчные.
0	0	0	Неконкретное слово ищется в словаре с переводом прописных букв в строчные.
0	0	да	Неконкретное слово ищется в словаре без перевода прописных букв в строчные без учета ударения.
0	да	0	Неконкретное слово ищется в словаре с переводом прописных букв в строчные с учетом ударения.

Для определения неконкретных искомых слов разработаны управляющие символы (+,=,?,*,@,^,#) (Табл. 4). Примеры конструирования неконкретное слов приводятся в табл. 5.

Таблица 4 – Описание употребления возможных управляющих символов для построения маски искомого слова

Управляющие символы	Объяснение
+, =	Поиск с учетом ударения (только при нажатом Stressed Mask)
?	Любой один символ
*	Любое количество символов (0, 1, 2 … n)
@	Любой один гласный символ
^	Любой один согласный символ
#	Любой один символ не гласный и не согласный

Таблица 5 – Примеры построения примерных слов для поиска в словаре

Неконкретное слово	Объяснение
??+??	Найти все слова с 4-мя символами с ударением на 2-ю.
??+*	Найти все слова с минимум 2-мя символами с ударением на 2-ю.
@^@^@^+	Найти все слова с 6-ю символами, со строением слова «гласный, согласный, гласный, согласный, гласный, согласный» и с ударением на 3-ю гласную.
#	Найти все слова с любым количеством символов, среди которых должны быть символы не гласные и не согласные.
мам*	Найти все слова с любым количеством символов, но с началом на “мам”. Причем при нажатом Stressed mask выберутся слова с безударной первой гласной.
а=виа*	Найти все слова с любым количеством символов, но с началом на “авиа”. Выберутся слова с частичным ударением на первый символ “а”, но без ударения на “и” и с главным ударением на одну из гласных после символа “и”.
+=*	Найти все слова с любым количеством символов. Выберутся слова с главным ударением, предшествующим частичному ударению.

Если настройки выбраны и слово (конкретное или неконкретное) введено, то через команду “Go!” программа АВ обратится к грамматическому словарю и выведет количество найденных

ответов (“Result:”) и сами результаты поиска в четыре колонки: слово, слово с обозначенным ударением, тег, приоритет (Рис. 11а, б, в, г).

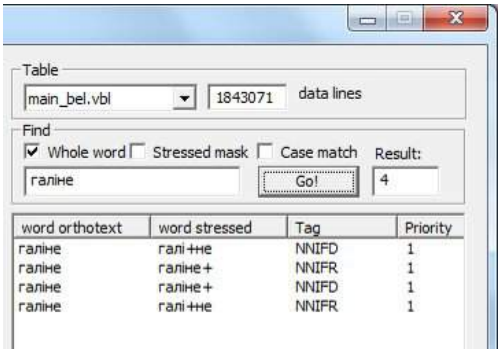


Рисунок 11а – Поиск в словаре конкретного слова для беларуского словаря

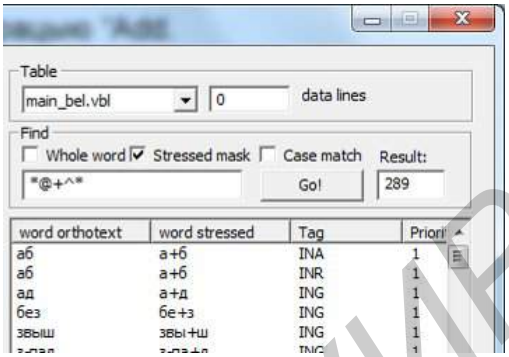


Рисунок 11б – Поиск в словаре неконкретного слова для беларуского словаря

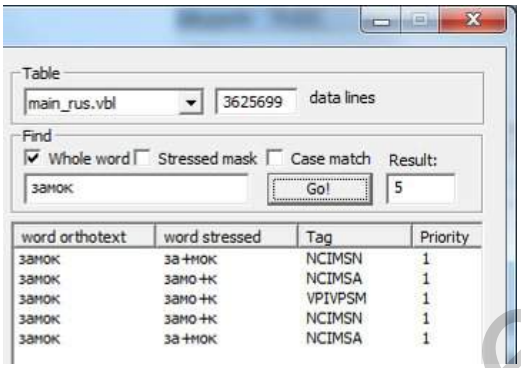


Рисунок 11в – Поиск в словаре конкретного слова для русского словаря

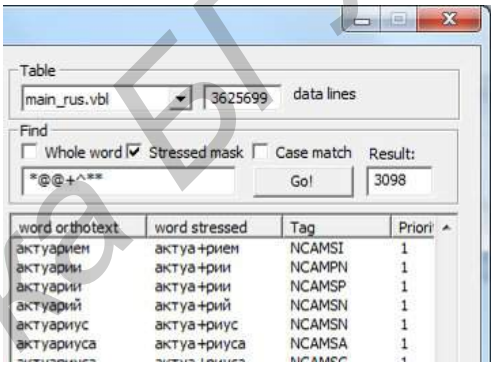


Рисунок 11г – Поиск в словаре неконкретного слова для русского словаря

Поиск слова может быть остановлен командой “Stop Action” (Рис. 12).

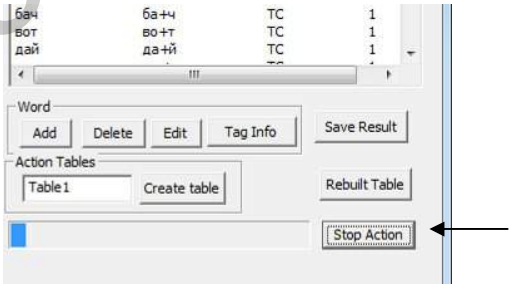


Рисунок 12 – Остановка поиска в словаре

Над любым выбранным словом в результате поиска может быть выполнена специальная операция (Рис. 13):

- Удалить (Delete)
- Редактировать (Edit)
- Показать расшифровку тега (Tag Info)

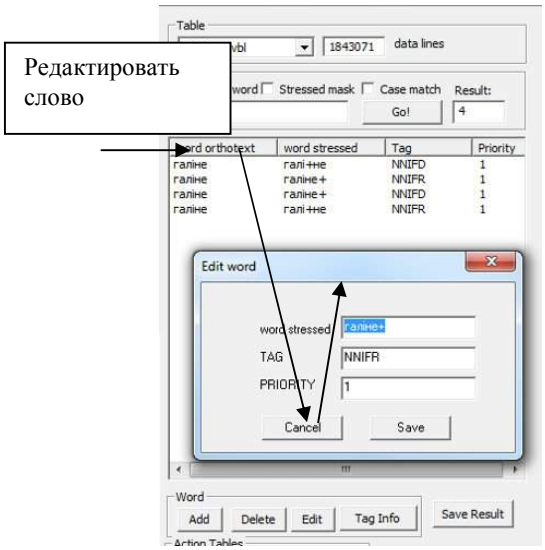


Рисунок 13 – Пример редактирования словоформы

Добавить слово в словарь можно через операцию “Add”.

В интерфейсе непосредственной работы со словарем есть кнопка перестроения основного файла словаря (например, main_bel.vbl) и хэша словаря (например, main_bel.hsh) –“Rebuilt” (Рис. 14). Ее можно использовать в двух случаях. Во время того, как слова удаляются из словаря, для быстроты работы ставится только отметка на слове, что оно удалено, а физически слово остается в файле *.vbl, чтобы уменьшить файл словаря. Также, когда происходит повреждение или случайное удаление файла *.hsh словаря, его можно обновить, перестроив словарь.

Рисунок 14 – Операция перестроения словаря “Rebuilt”

Чтобы создать новый словарь, нужно воспользоваться операцией “Create table” (Рис. 15). Словарь создается с написанным произвольным именем (по умолчанию Table1.vbl) и в него добавляется одно тестовое слово “ма+ма”.

(а) (б)

Рисунок 15 – Пример создания (а) и использования (б) нового словаря

Заключение

Таким образом, система редактирования и пополнения словарей речевого интерфейса для беларуского и русского языков позволяет пользователю улучшить словари для синтезатора речи находить новые и корректировать уже добавленные слова, обрабатывать слова-омографы и фиксировать нераспознанные выражения в тексте.

Важно отметить, что система может быть еще более доработана для удобства пользователей. Можно использовать различные подсвечивания цветом ненайденных слов и нераспознанных выражений в обработанном тексте. Также есть возможность разработать автоматическую систему подсказок ударений и ЛГК для новых слов по статистическим данным, которые можно брать из слов словаря, а пользователь будет только соглашаться с предложенными вариантами. В непосредственной работе со словарем можно добавить кнопку для прослушивания найденных слов, это позволит редактору словарей работать с меньшим визуальным напряжением.

Благодарности

Автор благодарен научному руководителю д.т.н. Лобанову Б.М. за содействие в проведении практической части и написании статьи, а также магистру филологических наук Денисюк С.А. за консультации в филологических вопросах и перевод статьи на русский язык.

Автор благодарен Беларускому республиканскому фонду фундаментальных исследований за поддержку исследований в рамках соглашения Ф10Р-006 по теме "Интеллектуальная модель синтеза выразительной речи на базе глубокого лингвистического анализа читаемого текста".

Библиографический список

[Бірыла, 1987] Слоўнік беларускай мовы: Арфаграфія. Арфаэпія. Акцэнтуацыя. Словазмяненне / Ін-т мовазнаўства імя Я.Коласа АН БССР; Пад рэд. М.В. Бірылы. – Мн.: БелСЭ, 1987. – 903 с.

[Лобанов и др., 2008] Лобанов, Б.М. Компьютерный синтез и клонирование речи / Б.М. Лобанов, Л.И. Цирульник // Минск: Беларуская наука, 2008. – 344 с.: ил.

[Житко и др., 2010] Житко В.А., Вяльцев В., Гецевич Ю.С., Розалиев В.Л, Естественно-языковой интерфейс вопросно-ответных систем [Работы данной конференции]

[Зализняк, 1980] Зализняк А.А. Грамматический словарь русского языка: Словоизменение. Ок. 1000000 слов. – 2-е изд. Стереотип. – М.: Рус. Яз., 1980. – 880 с.

[Lobanov et al., 2006] Lobanov, B., Tsirulnik, L. Development of multi-voice and multi-language TTS synthesizer (languages: Belarussian, Polish, Russian) // Speech and Computer: proceedings of the 11-th International conference SPECOM'2006, St. Petersburg, Russia, 25-29 June, 2006 / Institute of Informatics and Automation of RAS, Speech Informatics Group. – St.-Petersburg: Anatolia, 2006. – P. 274-283.