УДК 811.161.2'42:003.03

В. Г. Волошин, Н. В. Петлюченко

Проблемы предварительной обработки орфографического текста для синтеза украинской речи

В статье представлена информация относительно предварительной подготовки орфографического украинского текста для его дальнейшего преобразования в синтезированную речь. Основным этапом этой подготовки является составление словаря базовых лексем и словоформ с последующей расстановкой ударения и фонетической транскрипцией.

Ключевые слова: синтез речи, текстовый процессор, предварительная обработка, пофразовая обработка, пословная обработка, база данных слов и словоформ, фонемное транскрибирование

В основу практической реализации системы синтеза украинской речи по орфографическому тексту, разрабатываемой в ОНУ, положены алгоритмы аллофонного синтеза русской речи [4: 43–54]. Родственная близость двух славянских языков позволяет надеяться, что в целом структура алгоритмов будет сохранена, а изменения коснутся в основном специфики наполнения лингвистических баз данных и некоторых правил формирования речевого сигнала. Общая структура синтезатора русской речи [3: 101–105] включает четыре относительно самостоятельных процессора: текстовый, просодический, фонетический и акустический, реализующих последовательное преобразование орфографического текста в звучащую речь. В данной статье описываются особенности реализации применительно к синтезу украинской речи первого из них, текстового процессора.

Текстовый процессор предназначен для преобразования входного орфографического текста в размеченный фонемный текст. Под разметкой понимается разбиение текста на отдельные элементы в следующей иерархии: фонетический период, фраза, синтагма. Кроме того, процессор осуществляет: расстановку словесных ударений и интонационную маркировку синтагм. В общем виде текстовый процессор представляет собой совокупность трех основных блоков: предварительная обработка текста, пофразовая обработка текста, пословная обработка текста и совокупная база лингвистических данных и знаний (рис. 1).

Первый этап подготовки текста-документа осуществляется **блоком предварительной обработки текста**.

Назначение первого блока (рис. 2) состоит в предварительной обработке текста, в его нормализации, в приведении текста к каноническому виду.

Блок предварительной обработки текста выполняет следующие операции:

— операцию очистки текста от служебных знаков, не имеющих отношения к речи (знак переноса строки, табличные знаки и т. д.), что приводит текст, который виден на экране, в нормализованный орфографический текст;

— операцию преобразования всевозможных сокращений и аббревиатур в линейный текст (например: сокращения "и т. д." преобразуется в "и так далее", аббревиатуры "СНГ" — в "эс эн гэ", "США" — в "сэ шэ а", "ФРГ" — в "эф эр гэ";

— операцию преобразования "число-числительное", т. е. преобразования цифр в их орфографическое представление (например: цифры "28453" преобразуется в числительные "двадцать восемь тысяч четыреста пятьдесят три". Чтобы синтезировать произношение любого числа, требуется менее сотни базовых слов, таких как "один", "одна", "два", "две", "три", ... "сто", "ста" и т. д.;

— операцию преобразования формул (математических, физических, химических и т. д.) в их орфографическое представление.

Основное назначение **блока пофразовой обработки текста** (рис. 3) состоит в его просодической разметке.

Вначале осуществляется членение текста на фонетические периоды, затем на фразы и, наконец, на синтагмы. Фонетическим периодом называется наибольший участок речи, который еди-

нообразно оформлен с точки зрения интонации и ритмики. Обычно он соответствует такому отрезку текста, который называется в орфографии "абзацем". Далее этот текст членится на фразы. Фразы чаще всего соответствуют предложениям или части сложного предложения. Более сложная задача — членение фразы на синтагмы (если это необходимо, т. к. фраза может состоять только из одной синтагмы). Предложения в тексте могут быть очень длинными, обычно человек читает их не на одном дыхании, а разделяя на какие-то элементы по 3–4 слова, после которых допускается некоторая дыхательная пауза.

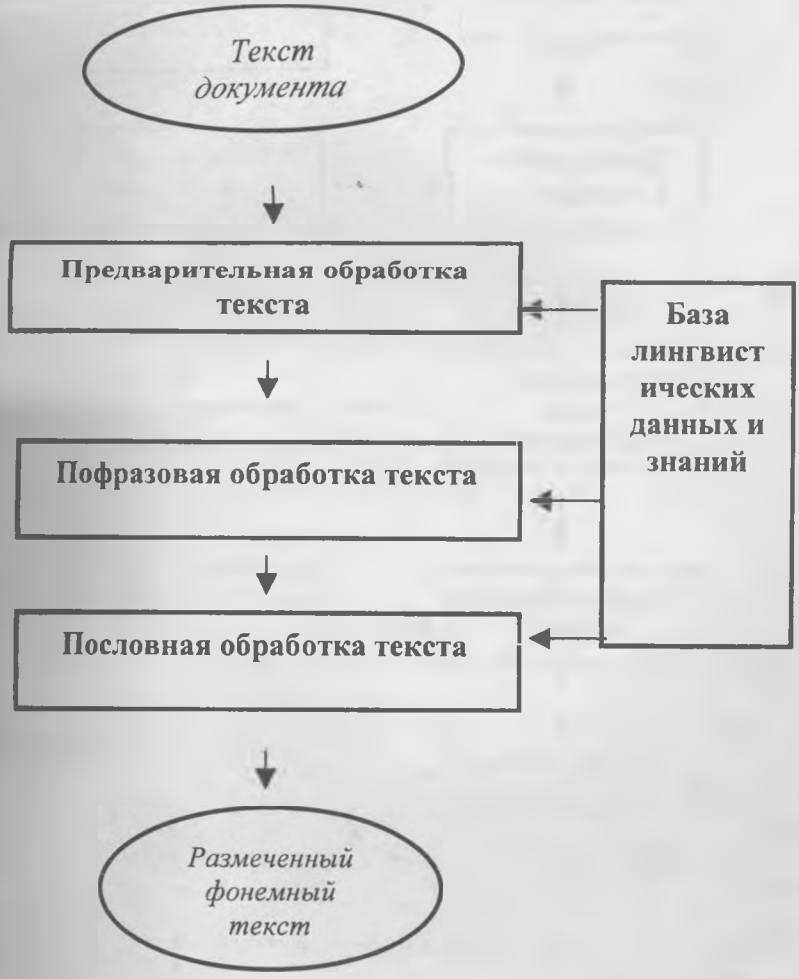

Рис. 1. Текстовый процессор

После членения текста на синтагмы эти синтагмы должны быть промаркированы фразовыми ударениями. После того, как промаркированы фразовые ударения, осуществляется интонационная разметка синтагм, т. е. исходя из того, какая синтагма является более или менее выраженной, где она находится во фразе, какой есть знак препинания, определяется интонационный тип синтагмы. Кроме интонационной разметки синтагм, необходимо установить длительность паузы, которая должна быть реализована после каждой синтагмы (паузация).

В результате работы блока пофразовой обработки текста получается просодически размеченный текст. В зависимости от того, как разбить фразу на синтагмы, звучание текста может быть самым разным и даже вообще изменить смысл предложения. Поэтому во всех этих блоках желательно использовать всю информацию, весь арсенал лингвистики: лексику (словарь), морфологию, синтаксис и семантику.

Рассмотрим конкретный пример превращения орфографического текста в просодически размеченный текст. Отрывок текста, использованный для иллюстрации, представляет собой типичный фонетический период, равный абзацу.

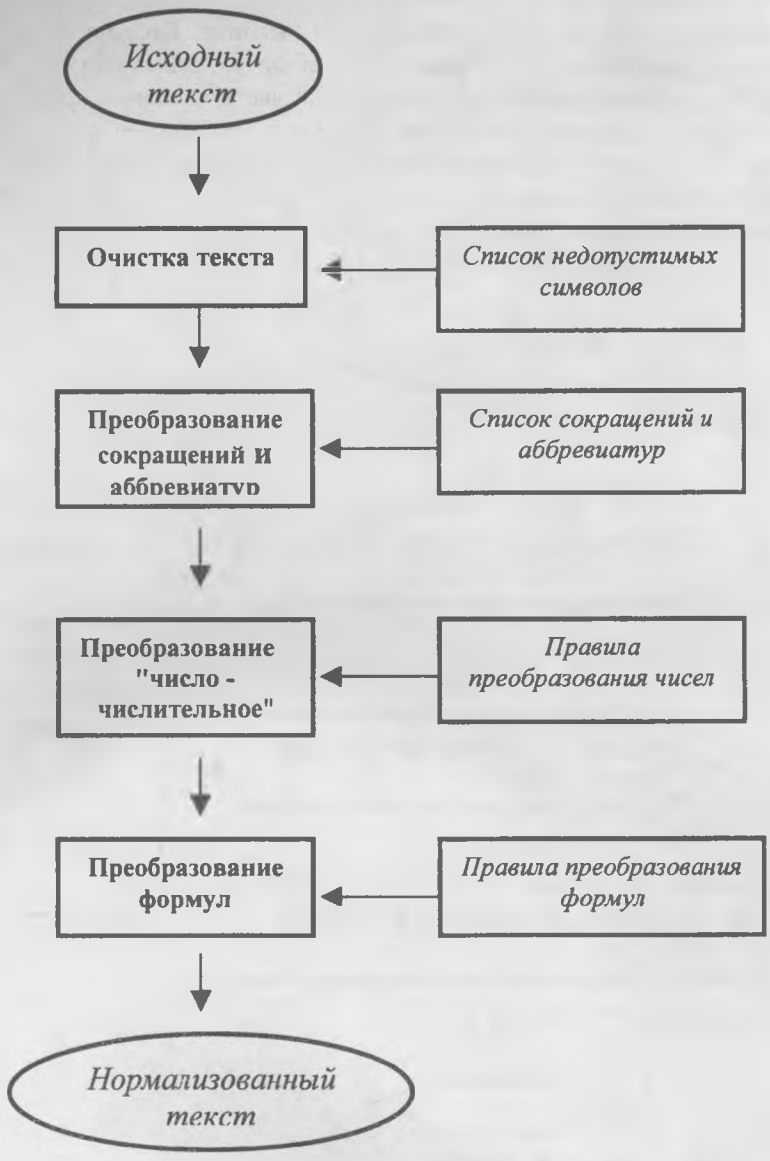

Рис. 2. Блок предварительной обработки текста

Исходный орфографический текст:

— Ви, як видно, ще не розумієте, що людину могли чекати друзі, а його запізнення на цілу добу руйнує всі плани і може викликати масу незручностей.

— Ах! Так справа була в цьому?

— От саме!

Анализируемый отрывок текста состоит из фраз разной длины. Первая фраза очень длинная и состоит из нескольких синтагм, вторая фраза состоит всего лишь из одного слова, третья и четвертая фразы — из одной синтагмы. Также эти фразы различаются интонационно: первая и четвертая — повествовательные или фразы с завершенной интонацией, вторая — восклицательная, третья — вопросительная.

Рассмотрим более подробно правила членения на синтагмы первой, самой длинной фразы.

Первым признаком границ между синтагмами являются знаки препинания. Без всякого риска конец синтагмы можно поставить также перед союзом "і". Граница синтагмы не должна стоять между синтаксически связанными словами, например, между определяемым и определяющим словом. Самые надежные критерии связанности слов — синтаксические правила. Но можно судить о границе синтагмы по более простым правилам, связанным с анализом частей речи. На-

пример, существительные и прилагательные, местоимения и существительные никогда нельзя расчленять, т.к. они жестко связаны друг с другом. Если же это существительное и глагол или два существительных, то они расчленяются.

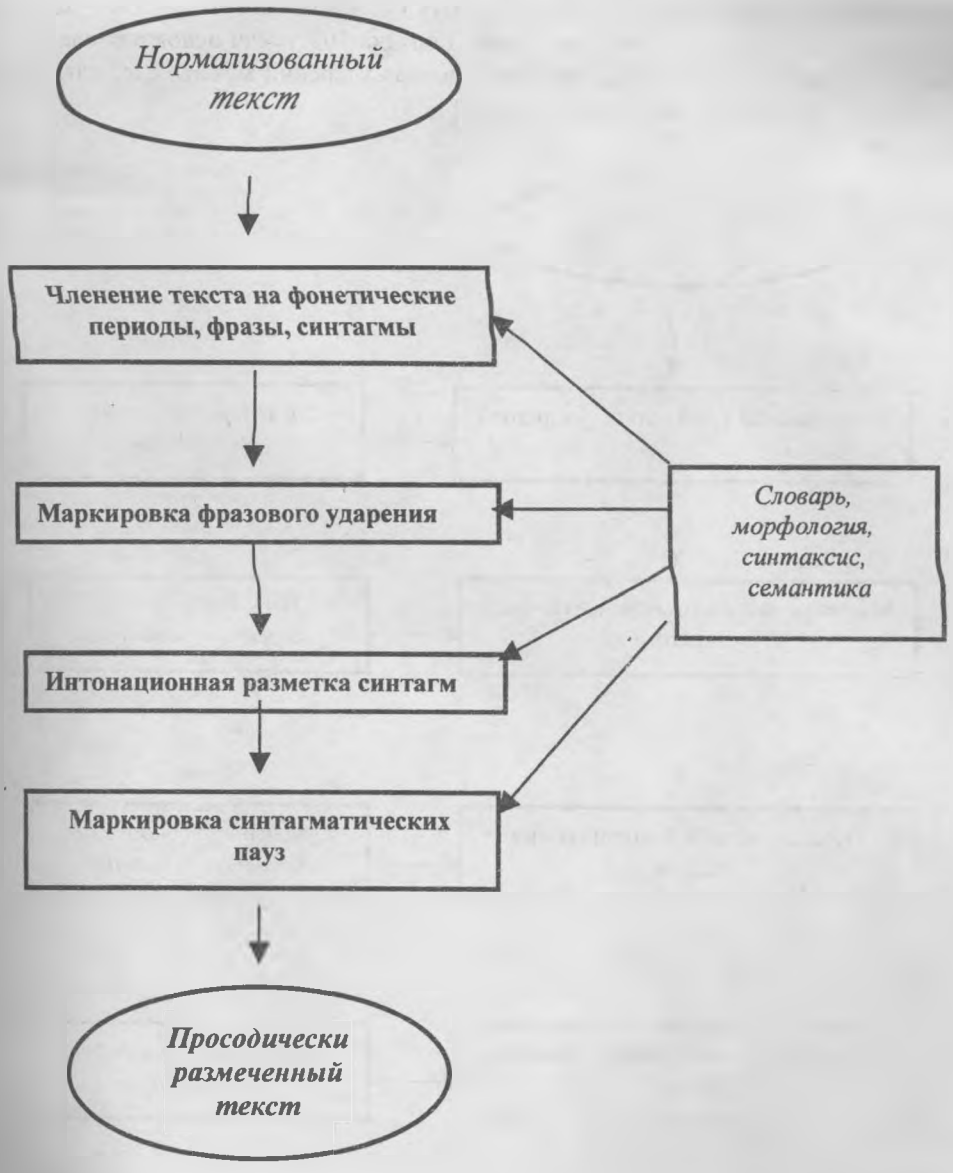

Рис. 3. Блок пофразовой обработки текста

В соответствии со сказанным получим следующую просодическую разметку текста:

— Ви, // як видно, // ще не розумієте, // що людину могли чекати друзі, // а його запізнення на цілу добу / руйнує всі плани / і може викликати масу незручностей. ///

— Ах! /// Так справа була в цьому? ///

— От саме! ///

Здесь знаки "/" обозначают конец синтагмы, а их количество — длительность синтагматической паузы.

Рассмотрим третий блок — **блок пословной обработки текста** (рис. 4).

Этот третий блок может уже не обращаться ко всей фразе, а только к каждому отдельному слову. Вначале осуществляется расстановка словесных ударений. Известно, что в украинском языке ударение свободное, т. е. оно может находиться на любом слоге, в отличие, например, от французского языка, где ударение всегда на последнем слоге слова, от чешского языка, где

ударение всегда на первом слоге, от польского языка, где ударение всегда на предпоследнем слоге. В украинском языке таких четких правил нет, поэтому, для того, чтобы проставить ударение, необходимо иметь словарь ударений. Это означает, что нужно иметь полный словарь украинского языка, если система претендует быть системой синтеза речи по тексту неограниченного словаря, т. е. нужно хранить в словаре порядка 100 тысяч основных словоформ, а также десятки их модификаций. Таким образом, словарь ударений может содержать более миллиона различных словоформ украинского языка [1; 3].

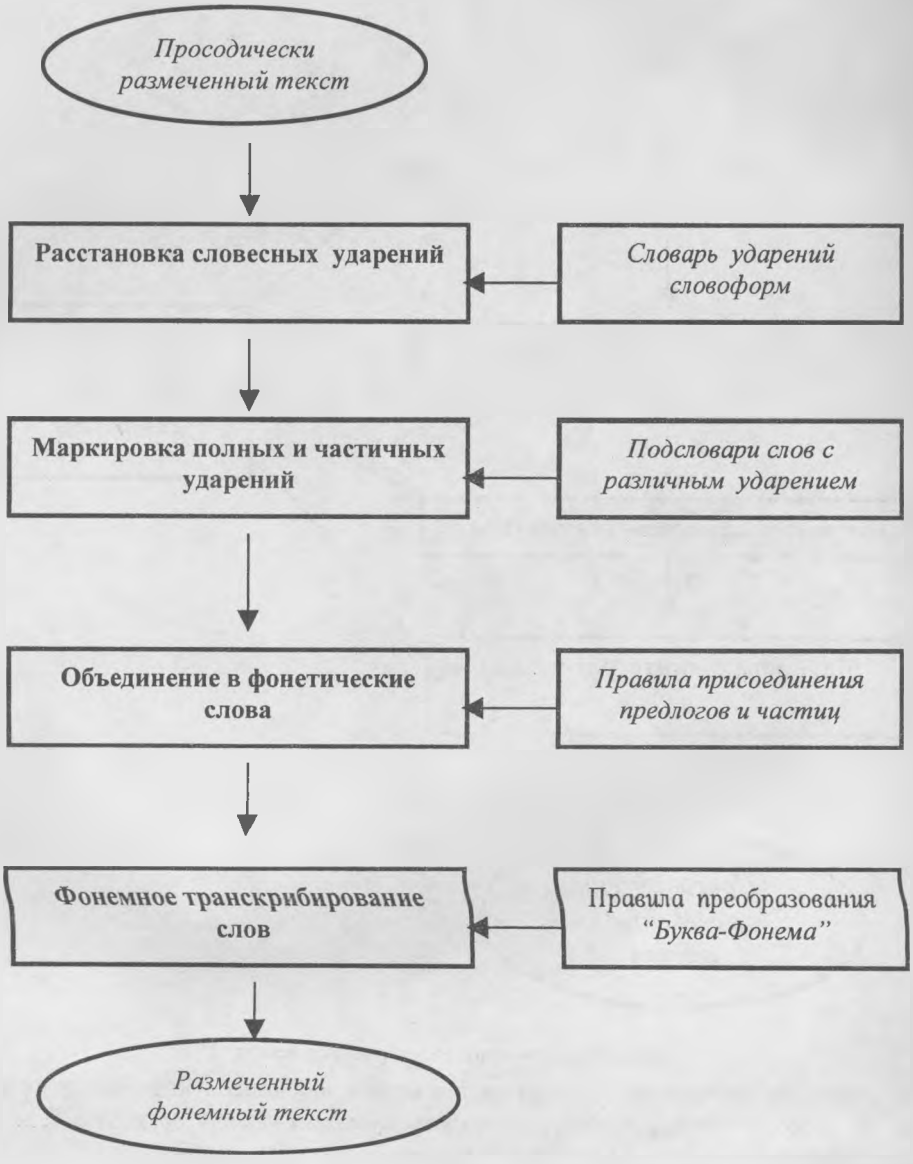

Рис. 4. Блок пословной обработки текста

Формирование базы данных слов и словоформ украинского языка основывается на фиксации лексем с обозначением ударения в цифровом виде. Слова могут располагаться как в алфавитном порядке, так и произвольно. Особые трудности при фиксировании слов возникают в следующих случаях:

1. В слова с двойным ударением (веснян'ий — весн'яний, комб'айнер — комбай'нер). В случае, если один из двух вариантов употребляется довольно редко, то этот вариант вообще не фиксируется.

2. В словах и словосочетаниях, ударение которых зависит от их семантики ('атлас — атл'ас, п'ора — пор'а).
3. В словах, дополнительные формы которых отличаются ударением, например, формы множественного числа (бал (банкет) — бал'и, бал'ів, бал (единица измерения) — б'али, б'алів).
4. В глаголах, в которых совершенный и несовершенный вид различается при помощи ударения (в'иводити — вив'одити, закл'икати — заклик'ати).
5. В словах с подвижным ударением. Если в окончании одного из косвенных падежей обозначается переход (смещение) ударения, то это является свидетельством того, что и в других падежах этот переход также будет происходить (баг'аж, -'у, — 'ем).
6. В словах, в которых производное употребление отличается от исходной формы (залік'овий — з'алік, перетр'имати — трим'ати).
7. В словах украинского языка, ударение которых отличается от ударения в их прямых лексических соответствиях в русском языке (верет'ено — веретен'о, кропив'а — крап'ива).

Дополнительные грамматические формы приводятся в таком виде, чтобы они отображали не только изменение ударения, но и чередование, выпадение, удвоение звуков, ассимиляцию и упрощение в группах согласных, являющиеся специфическими для украинского языка.

В сложных и сложносокращенных словах обозначается только основное ударение (високогірний, будь-що-б'удь), то же самое относится и к словосочетаниям (світ з'а очі, з'о сміху).

В связи с потребностью в полной акцентологической характеристике в словаре фиксируются все личные окончания глаголов (жити, живу, живеш, живе, живемо, живете, живуть), а также окончания редко используемых форм первого лица множественного числа (жив'ем, м'аєм, сид'им, сто'їм). В случае, когда личные окончания глаголов употребляются с двойным ударением, они фиксируются в такой последовательности: надпити, надіп'ю, надіп'єш, надіп'є, надіп'ємо, надіп'єте, надіп'ють, надіп'ю, надіп'єш, надіп'є, надіп'ємо, надіп'єте, надіп'ють.

В прошедшем времени глаголов указывается ударение как в мужском и женском роде (запр'іг, запрягл'а), так и в среднем роде и во множественном числе. В этих случаях ударение будет всегда на окончании (запрягл'о, запрягл'и).

После того, как будут проставлены ударения в каждом слове текста, эти ударения нужно промаркировать. Маркировка ударений необходима потому, что хотя большинство слов имеют полное (сильное) ударение, некоторые, например, местоимения, — только частичное (слабое) ударение, некоторые слова, такие как предлоги и частицы, могут вообще не иметь ударений. Поэтому, опираясь на тот же словарь, нужно промаркировать отдельные слова тем или иным типом ударений.

После маркировки ударений осуществляется процедура объединения слов в так называемые фонетические слова. Эта процедура заключается в объединении безударных слов со словами, у которых есть полное или частичное ударение, т. е. в объединении значащих слов со служебными: предлогами, частицами и союзами.

Последний этап — это **фонемное транскрибирование**. Оно поддерживается своими правилами. Правила транскрибирования иначе называются правилами преобразования "буква — фонема". При оценке правил преобразования букв в звуки необходимо составить список слов, которые по этим правилам будут иметь неправильное произношение и должны быть представлены в виде словаря исключений. В словарь исключений вносятся и слова-термины.

Имена собственные представляют особую проблему, поскольку их произношение часто определяется языком, лежащим в основе их правописания.

Ниже приводится пример преобразования рассмотренного ранее орфографического текста в размеченный фонемный текст:

— Ви,+ // йа-к в'и+дно, // шче- н'е розум+ієте, // шчо л'удину могли ч'екати дру+зі, // а його дал'+знен'ja на ц'ілу добу / рујнує вс'і плани / і може виклика+ти масу незру-чностеј. ///

— А+х! /// Так спра-ва була в ц'о+му? ///

— О-т са+ме! ///

Здесь знак (+) означает полное ударение, знак (–) — частичное, а знак (') после согласного означает его мягкость.

Фонетическая запись транскрипции текста далее оформляется наложением подходящего просодического контура для данного типа предложения на основании синтаксического анализа для разрешения некоторых фонетических неоднозначностей.

1. *Ганич Д. І., Олійник І. С.* Російсько-український і українсько-російський словник. — Харків, 1996.
2. *Головащук С. І.* Складні випадки наголошення. — Київ, 1995.
3. *Киселев В. В., Левковская Т. В. , Лобанов Б. М. , Хейдоров И. Э.* Синтезатор персонализированной речи по тексту "Лобано-Фон -2000" // 100 лет экспериментальной фонетике в России: Сб. научн. трудов междунар. конф. — Санкт-Петербург, 2001.
4. *Lobanov B. M.* Allophonic text-to-speech synthesizer: general structure and description // Автоматическое распознавание и синтез речи: Сб. научных трудов. — Минск, 2000.

В. Г. Волошин, Н. В. Петлюченко

ПРОБЛЕМИ ПОПЕРЕДНЬОЇ ОБРОБКИ ОРФОГРАФІЧНОГО ТЕКСТУ ДЛЯ СИНТЕЗУ УКРАЇНСЬКОГО МОВЛЕННЯ

У статті міститься деяка інформація відносно попередньої підготовки орфографічного українського тексту для його подальшого перетворення у синтезоване мовлення. Основним кроком цієї підготовки є формування словника базових лексем та словоформ з наступним визначенням наголосу у слові і його фонетичною транскрипцією.

Ключові слова: синтез мовлення, текстовий процесор, попередня обробка, пофразова обробка, пословна обробка, словник базових лексем та словоформ, фонетична транскрипція.

V. G. Voloshin, N. V. Petluchenko

PRELIMINARY PREPARATION OF ORTHOGRAPHIC TEXT FOR UKRAINIAN SYNTHETIC SPEECH

This paper gives some information regarding preliminary preparation of orthographic Ukrainian text for transforming into synthetic speech. The main step of this preparation is forming of basic lexemes and word forms vocabulary followed by identification of word accent and phonetic transcription.

Keywords: synthetic speech, text processor, preliminary preparation, phrase preparation, word preparation, basic lexemes and wordforms, phonetic transcription.