



Международная конференция по компьютерной лингвистике

[ENG](#) [Home](#) [Email](#) [RSS](#)

[Диалог 2014](#)[Предыдущие конференции](#)[Сборник материалов](#)[Форум](#)[Соревнование парсеров](#)[О конференции](#)[2013](#)[2012](#)[2011](#)[2010](#)[2009](#)[2008](#)[2007](#)[2006](#)[Архив](#)[2002](#)[2001](#)[2000](#)[1996](#)

[Главная](#) > Сборник «Компьютерная лингвистика и интеллектуальные технологии» > Архив > Труды Международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям

Спонсорам Диалога

Подписка на новости

Контакты

Следите за новостями

Найдите нас на Facebook



Dialogue

Нравится

255 пользователям нравится Dialogue.



Социальный плагин Facebook

20.12.2013 На нашем сайте опубликовано Первое информационное письмо. [Подробнее](#)

01.11.2013 Проект по созданию электронной версии собрания сочинений Л. Н. Толстого «Весь Толстой в один клик» ищет волонтеров. [Подробнее](#)

08.10.2013 Открыта подача предварительных заявок на конференцию «Диалог 2014». [Подробнее](#)

08.07.2013 Большое спасибо за фотографии с конференции «Диалог 2013» Савиной Настасье, Гиляровой Ксении, Новицкому Валерию, Сичинаве Дмитрию и Леонтьевой Анне. [Подробнее](#)

17.06.2013 Новые публикации о конференции «Диалог 2013». [Подробнее](#)

Проблемы предварительной обработки орфографического текста для синтеза украинской речи

Волошин В.Г., Петлюченко Н.В., Лобанов Б.М.

В основу практической реализации системы синтеза украинской речи по орфографическому тексту, разрабатываемой в ОГУ, положены алгоритмы аллофонного синтеза русской речи [1]. 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 [2] включает четыре относительно самостоятельных процессоров: текстового, просодического, фонетического и акустического, реализующих последовательное преобразование орфографического текста в звучащую речь. В данном докладе 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).



Рис. 1. Текстовый процессор

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";



Рис.2 Блок предварительной обработки текста

- операцию преобразования "число-числительное", т.е. преобразования цифр в их орфографическое представление (например: цифры "28453" преобразуются 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.



Рис.3 Блок пофразовой обработки текста.

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

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.

Исходный орфографический текст:

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

-Ах! Так справа була в цому?

-От same!

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 связанныности слов - синтаксические правила. Но можно судить о границе синтагмы по более простым правилам, связанным с анализом частей речи. Например, 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.

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 / розбудовує всі плани / і може викликати масу незручностей.///

-Ах!// Так справа була в цому?//

-От same!//

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

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



Рис.4 Блок пословной обработки текста

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 [2; 3].

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'й – весн'яний, комбайнер – комбайнер). В случае, если один из двух вариантов употребляется довольно редко, то в тогда этот вариант вообще не фиксируется.
2. В словах и словосочетаниях, ударение которых зависит от их семантики ('атлас – атл'ас, п'ора – пор'a);
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 (заліковий – 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, крапив'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 (високог'ірний, будь-що-б'удь), то же самое относится и к словосочетаниям (світ з'a очі, з'o сміху).

В связи с потребностью в полной акцентологической характеристики в словаре

фиксируются все личные окончания глаголов (жити, живу, дживеш, живе, живемо, живете, живутъ), а также окончания редко используемых форм первого лица множественного числа (жив'ем, м'аэм, сид'им, сто'им). В случае, когда личные окончания глаголов употребляются с двойным ударением, они фиксируются в такой последовательности: надпiti, надіп'ю, надіп'еш, надіп'є, надіп'ємо, надіп'єте, надіп'ють, надіп'ю, надіп'еш, надіп'є, надіп'ємо, надіп'єте, надіп'ють.

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

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.

Нравится 0 Tweet 0 g+1 0

Мне нравится +1



По всем вопросам обращайтесь по адресу: Secretary@dialog-21.ru