МІНІСТЭРСТВА АДУКАЦЫІ РЭСПУБЛІКІ БЕЛАРУСЬ

БЕЛАРУСКІ ДЗЯРЖАЎНЫ УНІВЕРСЫТЭТ

ФІЛАЛАГІЧНЫ ФАКУЛЬТЭТ

КАФЕДРА ПРЫКЛАДНОЙ ЛІНГВІСТЫКІ

Анатацыя да магістэрскай дысертацыі

ПАБУДОВА ЛІНГВІСТЫЧНЫХ РЭСУРСАЎ СІНТЭЗАТАРА МАЎЛЕННЯ ПА ТЭКСЦЕ ДЛЯ АГУЧВАННЯ ДАДЗЕНЫХ НАНА- І ПІКАСПАДАРОЖНІКАЎ

Барадзіна Юлія Станіславаўна

Кіраўнік: Гецэвіч Юрый Станіслававіч

Спецыяльнасць: Інавацыі ў навучанні мовам як замежным

2015

РЭФЕРАТ

Барадзіна Юлія Станіславаўна

Пабудова лінгвістычных рэсурсаў сінтэзатара маўлення па тэксце для агучвання дадзеных нана- і пікаспадарожнікаў

Склад дыпломнай працы: 63 старонкі, 20 малюнкаў, 4 табліцы, 41 крыніца, 2 дадаткі.
Ключавыя словы: колькасныя выразы з адзінкамі вымярэння, КВАВ, навукова-тэхнічны тэкст, тэлеметрыя, лінгвістычны рэсурс, сінтэз маўлення па тэксце, корпус, канкарданс, NooJ, руская як замежная, беларуская як замежная.

У працы апісваецца працэс стварэння ўніверсальных лінгвістычных рэсурсаў па апрацоўцы колькасных выразаў з адзінкамі вымярэння (напрыклад, *756 км/г, 67,6 кПа, 1368 В*) на рускай і беларускай мовах. Апрацаваны тэкст уяўляе сабой выразы тыпу *семсот пяцьдзесят шэсць кіламетраў у гадзіну*, *шэсцьдзесят сем цэлых шэсць дзесятых кілапаскаля* або*тысяча трыста восемдзесят шэсць вольт*.

Распрацаваныя рэсурсы мы завем універсальнымі, таму што яны ахопліваюць вялікі лікавы дыяпазон, мноства структурных схем лічэбнікаў і маюць каля 120 адзінак вымярэння. Акрамя таго, мы прапануем яшчэ і два спосабы прыкладання распрацаваных рэсурсаў: агучванне пры дапамозе сінтэзатара маўлення па тэксце дадзеных спадарожнікаў, якія таксама называюцца тэлеметрыяй; і выкладанне тэмы лічэбнікаў і адзінак вымярэння на занятках па рускай або беларускай мовах як замежных у тэхнічных аўдыторыях (ці рускай/беларускай ў спецыяльных мэтах).

Актуальнасць работы абумоўліваецца недастатковай распрацаванасцю падобнага роду рэсурсаў для беларускай і рускай моў, у той час як іх стварэнне дазволіць палепшыць якасць сінтэзу маўлення па тэксце, якасць інфармацыйнага пошуку, можа служыць базай для стварэння даведачнага матэрыялу па скланенню рускіх і беларускіх лічэбнікаў для рэдактараў, карэктараў, тэхнічных пісьменнікаў і інш .

Для дасягнення пастаўленых задач выкарыстоўваюцца метады корпуснай лінгвістыкі, інфармацыйнага пошуку і колькасныя метады статыстыкі, а таксама прыватныя метады такой галіны ведаў як камп’ютарная лінгвістыка. У якасці аб'екта даследавання былі выбраны спалучэнні колькасных лічэбнікаў і адзінак вымярэння пры іх, якія мы таксама завем колькаснымі выразамі з адзінкамі вымярэння (КВАВ).

Усе лінгвістычныя рэсурсы рэалізаваны пры дапамозе камп'ютэрна-лінгвістычнай праграмы NooJ. Вынікі даследавання будуць прапанаваны для

ўкаранення ў лабараторыі сінтэзу і распазнавання маўлення АІПІ НАН Беларусі.

РЕФЕРАТ

Бородина Юлия Станиславовна

Построение лингвистических ресурсов синтезатора речи по тексту для озвучивания данных нано- и пикоспутников

Состав магистерской работы: 63 страницы, 20 рисунков, 4 таблицы, 41 источник, 2 приложения.

Ключевые слова: количественные выражения с единицами измерения, КВЕИ, научно-технический текст, телеметрия, лингвистический ресурс, синтез речи по тексту, корпус, конкорданс, NooJ, русский как иностранный, белорусский как иностранный.

В работе описывается процесс создания универсальных лингвистических ресурсов по обработке количественных выражений с единицами измерения (например, *756 км/ч, 67,6 кПа, 1368 В*) на русском и белорусском языках. Обработанный текст представляет собой выражения типа *семсот пятьдесят шесть километров в час*, *шестьдесят семь целых шесть десятых килопаскаля* или *тысяча триста шестьдесят восемь вольт*. Разрабатываемые ресурсы мы зовем универсальными, потому что они охватывают большой числовой диапазон, множество структурных схем числительных и имеют порядка 120 единиц измерения. Кроме того, мы предлагаем еще и два способа приложения разработанных ресурсов: озвучивание при помощи синтезатора речи по тексту данных спутников, которые также называются телеметрией; и преподавание темы числительных и единиц измерения на уроках русского или белорусского языков как иностранных в технических аудиториях (или же русского/белорусского в специальных целях). Актуальность работы обусловливается недостаточной разработанностью подобного рода ресурсов белорусском и русском языках, в то время как их создание позволит улучшить качество синтеза речи по тексту, качество информационного поиска, может служить базой для создания справочного материала по склонению русских и белорусских числительных для редакторов, корректоров, технических писателей и др.

Для достижения поставленных задач используются методы корпусной лингвистики, информационного поиска и количественные методы статистики, частные методы такой области знаний как компьютерная лингвистика. В качестве объекта исследования были выбраны сочетания количественных числительных и единиц измерения при них, которые мы

также называем количественными выражениями с единицами измерения (КВЕИ).

Все лингвистические ресурсы реализованы при помощи компьютерно-лингвистической программы NooJ. Результаты исследования будут предложены для внедрения в лаборатории синтеза и распознавания речи ОИПИ НАН Беларуси.

ABSTRACT

Julia Borodina

Creation of linguistic resources of text-to-speech synthesizer for voicing nano- and picosatellites data

Content of the diploma paper: 63 pages, 20 figures, 4 tables, 41 source, 2 appendices.

Keywords: quantitative expressions with measurement units, QEMU, scientific and technical texts, telemetry, linguistic resource, text-to-speech synthesis, corpus, concordance, NooJ, Russian as a foreign language, Belarusian as a foreign language

The paper describes the process of building universal linguistic resources for the processing of quantitative expressions with measurement units (e.g. *756 km/h, 67.6 kPa, 1368 V*).for Russian and Belarusian language. Processed text looks like fully unwrapped phrase, e.g. *seven hundred fifty-six kilometer per hour, sixty-seven point six kilopascal, one thousand three hundred and sixty-eight volt*.

We call our resources universal because of the large number diapason, vast variety of measurement units and lots of structure schemes for numerals. Additionally, we propose two application options for those linguistic resources: one of them is the processing of satellite data, which is also known as telemetry; the other one describes usage of linguistic resources to learn the topic of numerals and measurement units during lessons of Russian as a foreign language or Belarusian as a foreign language.

Relevance of the work is caused by inadequate development of linguistic resources of this kind in Belarusian and Russian languages, while their creation can improve the quality of text-to-speech synthesis, the quality of information retrieval, can serve as a basis for creating a reference system of Russian and Belarusian numerals for editors, proofreaders, technical writers and others.

To achieve these objectives we using methods of corpus linguistics, information retrieval and quantitative methods of statistics, as well as private methods of such discipline as computational linguistics. As the object of study were chosen combination of numerals and measurement units attached to them, which we also call quantitative expressions with measurement units (QEMU).

All the linguistic resources were realized in linguistic development software called NooJ. The results will be proposed for implementation in the Speech Synthesis and Recognition Laboratory of UIIP NASB.