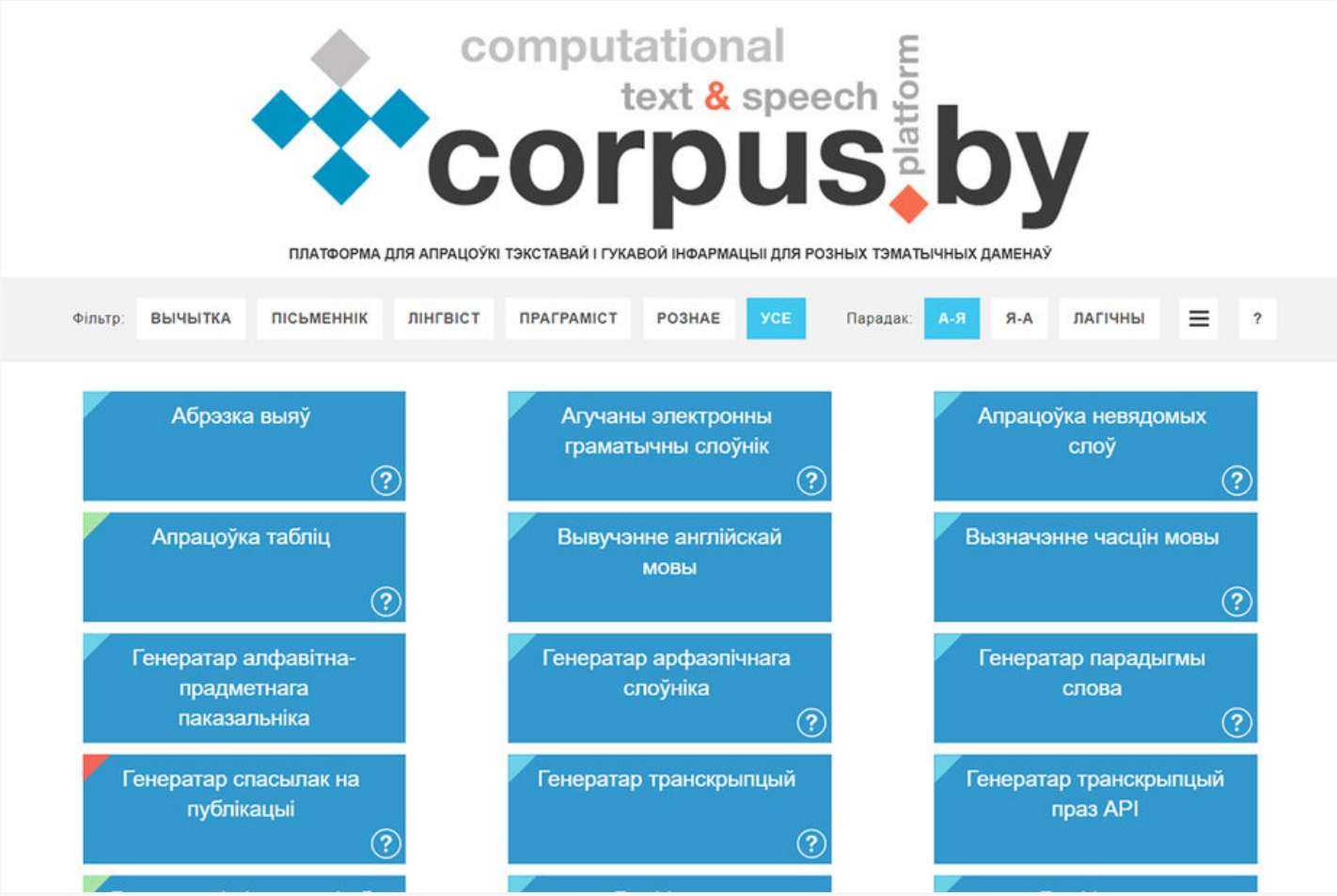
30.07.2017 / 20:57

Захоўвайце спасылку — сайт, які спросціць жыццё ўсім беларускамоўным

Мы шмат гаварылі пра першы беларускі арфаэпічны слоўнік, але не распавялі самае галоўнае — дзякуючы чаму ён з'явіўся.

Большасць працы за людзей выканала машына, якая аўтаматычна згенеравала транскрыпцыю 117 тысяч беларускіх слоў з дакладнасцю 98%. А вось распрацавалі машыну спецыялісты Лабараторыі распазнавання і сінтэзу маўлення Аб'яднанага інстытута праблем інфарматыкі Нацыянальнай Акадэміі навук.



Ужо даўно працуе цалкам бясплатная анлайн-платформа для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў Corpus.by. На ёй сабраныя дзясяткі сэрвісаў, якія дапамагаюць у вывучэнні беларускай мовы і не толькі. Стварылі яе супрацоўнікі лабараторыі. Пяць гадоў таму тут было ўсяго тры сэрвісы. Затое сёння сэрвісаў больш за 40. На распрацоўку некаторых пайшло паўдня, на іншыя — месяцы і гады.

Цяпер тут шмат магчымасцяў: хочаце — генеруйце транскрыпцыі, хочаце — запускайце сэрвіс «гаворачая галава» і глядзіце на чалавека, які агучвае тое, што вы папросіце. Можна праверыць правапіс, падзяліць словы на склады, упарадкаваць словы па алфавіце ці (каб вы ўжо дакладна ўпэўніліся ў разнастайнасці рэсурсу) сканвертаваць тэкст у код Морзэ.

Асабліва карысны сэрвіс — «агучаны электронны граматычны слоўнік». Тут можна праверыць напісанне любога беларускага слова па ўсіх наяўных слоўніках і праслухаць, як яно гучыць у выкананні сінтэзатара маўлення.

Асцярожна — сайт засмоктвае!

Для зручнасці сэрвісы падзеленыя па секцыях: «Вычытка», «Пісьменнік», «Лінгвіст», «Праграміст» і «Рознае». Плануецца, што пазней з'явяцца асобныя секцыі для медыкаў, бібліятэкараў і фізікаў. То бок для кожнай прафесіі прадугледжаны свой набор інструментаў.

Уверх

Як Corpus.by звязаны з арфаэпічным слоўнікам

Калі ў Лабараторыю распазнавання і сінтэзу маўлення звярнулася ўкладальніца арфаэпічнага слоўніка Валянціна Русак з просьбай дапамагчы ў распрацоўцы фаліянта, праграмісты прыдумалі тэхнічнае рашэнне, дзякуючы якому ўдалося зэканоміць не адзін год працы.

«Усё пачалося з таго, што наша лабараторыя распрацавала аўтаматычную сістэму сінтэзу маўлення і выклала яе ў інтэрнэт для вольнага карыстання, — распавядае загадчык лабараторыі Юрась Гецэвіч. — Згенераванае маўленне не самае натуральнае, але разабраць словы можна. Сінтэзатар звычайна выкарыстоўваецца для таго, каб паказаць студэнтам і выкладчыкам, як у прынцыпе любы беларускамоўны тэкст можа ператварыцца з паслядоўнасці электронных сімвалаў спачатку ў арфаграфічны тэкст, а потым — у фанетычны, які ў выніку прагаворвае машына. Гэта вельмі важна разумець, каб будаваць чалавека-машынны інтэрфейс».

Юрась Гецэвіч. Фота movananova.by

Генератар працуе не з запісанымі словамі, а з запісанымі асобнымі гукамі.

«Калі я рабіў лабараторную ва ўніверсітэце ў межах курса, які вёў Юрый Гецэвіч, мы працавалі недзе з 80 гукамі, — прыгадвае малодшы навуковы супрацоўнік лабараторыі Станіслаў Лысы. — Гэта было цікава, мы сінтэзавалі розныя тэксты, але не маглі зразумець, што хто насінтэзаваў. Тады падыходзіў Юрый і казаў: «Ага, ну, гэта з «Каласоў»» [»Каласы пад сярпом тваім» Караткевіча — НН]. Ён быў ужо спрактыкаваны і добра разумеў машыну. Канечне, 80 гукаў — гэта адно. А тысячы, якія мы маем цяпер, — зусім іншае».

Прынцып работы сінтэзатара маўлення пасля і выкарысталі для аўтаматычнай генерацыі транскрыпцый слоў. Станіслаў Лысы стварыў для гэтага асобны сэрвіс — «Генератар арфаэпічнага слоўніка». Спачатку яго тэставалі лінгвісты, якія вышуквалі памылкі і перадавалі іх у лабараторыю на выпраўленне. У выніку сэрвіс навучыўся генераваць транскрыпцыі амаль бездакорна.

«Нам удалося дамагчыся гэтых 98% дзякуючы некалькім сотням правілаў, якія прапісалі Барыс Лабанаў, Лілія Цырульнік, Дзмітрый Пакладок і скарэктавалі Алена Гюнтар, Яўгенія Зяноўка, Юрась Гецэвіч і я. Прабачце, калі не ўсіх распрацоўшчыкаў правілаў — супрацоўнікаў нашай лабараторыі — узгадаў. Чым больш вузкае правіла, тым складаней яго дадаць. Узяць тое ж «г» выбухное. Давялося ламаць галаву на тым, як яго ўключыць, каб не паламаць усё астатняе», — дадае Станіслаў Лысы.

Станіслаў Лысы.

Адзін з самых добра распрацаваных сэрвісаў — «Праверка правапісу». Праграма праганяе тэкст і па беларускіх слоўніках, і па замежных. Дарэчы, карыстальнік можа стварыць і дадаць свой уласны слоўнік, якім іншыя пры жаданні таксама могуць карыстацца.

«Сэрвіс пакуль што не праяўляе інтэлект, не шукае сэнсавыя памылкі, а звяраецца з усімі тымі базамі, што ў нас ёсць. Напрыклад, калі ў слове ёсць памылка, але праз яе ўтвараецца новае слова, якое існуе, такую памылку сэрвіс не ўбачыць. Калі ж слова не сустракаецца ні ў адным слоўніку, праграма абавязкова пакажа гэта, — тлумачыць Станіслаў Лысы.

— Бывае, чалавек устаўляе лацінскую літару «і» замест беларускай, і потым ніводная камп'ютарная праграма гэтае слова не разумее. Аднойчы мы нават адшукалі слова «арахіс», напісанае цалкам англійскімі літарамі. Праграма дае магчымасць убачыць усе змяшаныя напісанні».

Станіслаў дадае, што сэрвісы Corpus.by увесь час паляпшаюцца.

«Мы адразу бачым актыўнасць карыстальніка і накіроўваем высілкі ў развіццё таго, што яму патрэбна», — тлумачыць Станіслаў Лысы.

Супрацоўнікі лабараторыі заклікаюць карыстальнікаў заходзіць на інтэрнэт-рэсурс Corpus.by і знаёміцца з сэрвісамі, амаль кожны з якіх мае падрабязнае апісанне. А таксама ўключацца ў іх удасканаленне і паведамляць пра тыя рэчы, якія можна палепшыць, а таксама памылкі, што варта выправіць.

Настасся Роўда