

Òåçàóðóñ ïî êîðíóñíé ëèíãâèñòèå

Àâòîð Àèêòîð Ìàäéïàè÷ Çàôàðïà
02.09.2010 ã.
Ìñëåäíå ïáíâëåíèå 03.09.2010 ã.

The paper is devoted to the construction of the Russian thesaurus on Corpus Linguistics. Linguistic resource involved in research is the Russian corpus on Corpus Linguistics developed in St.-Petersburg State University and Institute of Linguistic Studies and different vocabularies. Semi-automatic terminology extraction is performed with the help of linguistic and statistical tools which allow to generate lists of single-word and multi-word terms supplied with frequency data and lexical-syntactic patterns. Lexical-syntactic patterns are used in the analysis of contexts which contain definitions of terms, expose interrelations between terms, provide their synonyms, translation equivalents, etc.

Â iáñòiýùåå áðâðaiý â iáðæñòðe êiðiðiññíé ëeíââðeñòðeé èòñòðoñòðåðåð -ðâðòðæý ñèñòðiað iáð-ú-ííé ðâððiðiññíéñæð. Êiðiðiññíay ëeíââðeñòðiñòðiâíéý è èññiñëüçíââðeý ëeíââðeñòð-ðâññéø ëiðiðiññâ ñ èññiñëüçíââðeâí ëiññüþðoðiñûð òððoñííñæð. Íñä iáçââðeâí ëeíââðeñòð-ðâññéø èâðæä è iððââðaíacíá-ðâññûð æðy ðâððâðeý ëiññüþðoðiñûð ëeíââðeñòð-ðâññéø çâðâð. Êiðiðiññíay ëeíââðeñòðeé ìáððâðñâ-ðâíèé çâðâð ðâððiðiññæð. Áñ-íâðâðuñ, ýðiñ añðañðâðâðiñ, ó-ðèðùñâðy åâð iáððâðiñâð ðâððiññiæðâðeâ. Áñ-âðiðuñ, ëiðiðiññíay ëeíââðeñòðeé ëaâð iððâðeñûð ñèñéâðâðåðuñý â iáððâðo ñâðeññâðiñ ÿçûñêà. Þiýðiññí, ðâððâðâðo ñâðuñâðy òâðcâðoðoñ ïí ðâððiðiññíé ëeíââðeñòðeéâ, añðoññiñòðâðâðiñ ñâðâðiññæð. Ëo àíâðeññæð yéâðâðeâðiñ. Á iáñòiýùåå áðâðaiý ðâððeñ ðâðcâðoðoñíâ ìáñòðaññòðåðåð.

Êðóã äâèñèè, ìòíñjùåéñý ê ñôåðå êîðiöñíé èëíââèñòèè è ïïäéâæàùåé ôîðìàèèçàöè, î-âð-âí èïåþùèèñý ýöèéëïïäæ-âñèè ñïäæíú íàïðåâéäíÿ. Íïòíó î ðè ðàçðåáíòéå òàçåóðóñà íù èññïäèì èç íàëè÷èÿ ðÿäà ïññáéè, êîòíðûå íäóò áûòü èñíïëüçâàíû, à Èëíââèñòè-âñèè 2005, ßçûêíçíáíèå 2007 è äð.].

Đàcòàáòái à lõäöïà è iññöötlàíèp òâçáóööñïà íðåäílåòíûö íáëæñòåé, íðåäílæäàþùèé áûäääéåíèå èéþ-ååûö iñíýòèé èç lïíæñòåà èç äåññöèiöötlàìè, èçâéå-åííûìè èç ó-åáíûö iññiaéè ii êïööñïíé eeíáâèñöèéå, à -åñòíñòè [Çàööåööñïà 2005]. Ià àòöötlï yóäïå iðåäílæäè öåéñòö (iðåäääå âñåäî, ióáéèéàöèé êííôåðåöèé ii êïööñïíé eeíáâèñöèéå: «Êïööñïàÿ eeíáâèñöèéà è eeíáâèñöè-åñéèå áåçû à – 2004, 2006, 2008» è äö.). Ià áåçå ýóïäi êïööñïà ñïçåâàöñý íåðåäàÿ áåðñèý òâçáóööñïà ii êïööñïíé eeíáâèñöèéå. Båäöi y õåðåéòåðèçóþùeo íáëæñòü «êïööñïàÿ eeíáâèñöèéà» è lòèè-åþùèöñý áûñïîé è íòöötlàòéåíñòüþ è -åñòíòíñòüþ. Äëÿ iññöötláíèÿ èäéñèéè, ó-èoùåâàþùèé èíäéè-íñíýòèéíúå ñõåäi iðåäílåòíûé íáëæñòè, iðåäñòåâäéåííúå à yéñiäðöiûö iñèñäàëýö. Äëÿ àåòöiàòèçèööñïà èíëëíéåöèé, iñííåûåâàþùèåñý íà èçâáñòíûö íåðåò àññòéèåöèé (MI, T-score, Log-Likelihood).

Øåçàóðóñ áóäåò áàçèðîàòüñý íà ïíòïéñàè èîðlöñííé èèíâèñòèêè, ðàçðàáàòûâàåííé íà êàôååðå ìàòåìàòè÷åñîé èèíâèñòèêè èííüþòåðñíé èèíâèñòèêè, ðàçðàáîòàíÿà ëíàññèåðñêå à ðàïèæò íðîâèòà ïíñïçääíèþ ëíðöåëà çíàíèé íí èííüþòåðñíé èèíâèñòèêè

Òåçàóðóñ áóëåò ïðåäñòàâéþò ñíáîé ýéåèòðííóþ áàçó äàííûõ, íáññá÷èâàþùóþ ïíüçíàòåðéÿì äíñòóï ê íàíó. Êðíà òíñâ, áóëåò ñííññívcu÷íûõ (ISO 5964-1985, ÁÍÑÒ Đ 7.24-2007) òåçàóðóñâ.

Eèòåðàòóðà

Áàðàíî Á.Í. Áâåäåíéå á ïðèéëàäíóþ ëèíâàèñòèêó. Ñåðèý "Íîâûé ëéíâàèñòè÷åñèé ó÷åáíèé". І.: Ýæòîðèä ÖÐÐÑ. 2001.

Áiðiâèéiâà Í.È., Çàãiðóëüêî þ.À., Çàãiðóëüêî Á.Á., Èíñíáíêî È.Ñ., Ñièéiâà Å.Å. Ðàçðàáîòèà iñðòàëà çiàiéé iñ èñiüþòåðíé èëíâà (ä.Äóáìà, Ðiññèÿ). –l.: ËÁÍÁÍ, 2008. –Ó.3. –Ñ.380-388.

Âèññðåäññâ Í.Â., ìèòðîòàññâ Í.Â. Ôíðìæüíàÿ ñíòïëñäèÿ èàéè èíñòðóíåíò ñèñòåíàòèçàöèè àáííúô à ðóññëíÿçû÷íî êíðíøñâ òåéñòñâ í ÑíáÁÓ, 2008. – C. 113-121.

Äåìüyíéïâ Å.Ç. Äíæëî-ðóññêèå òåðíèíû ïï iðèéëàäííé ëèíâèñòèêå è àâòííàòè÷åññéíé íåðåðàáíòèå òåðñòà. Äüi. 2. ìåòíäû àíàëèçà

Çàõõàðîâ Ä.Í. Êïðþóñíàÿ ÿéíââèñòèêà: Ó-åáíî-ìåòëè-÷åñîå ìñíàèå. – Ñíá.: ÑíáÅÓ, 2005. – 48 ñ.

Êèíââèñòè÷åñêèé ýíoèéëïääè÷åñêèé ñëîâàðü. Ì.: Ñîâ. Ýíoèéëïääèý, 1990.

Íèòðîòàíâà Í.À., Çàõàðîâ Á.Í. Àâòìàòèçèðîâàíûé àíàëèç òåðìèíïëäè â ðóññëîýçû÷íî êðïóñå òåêñòîâ íî êðïóñíé èëíââèñòèå «Äèàëíâ–2009». Í.: 2009.

Íèêèòèíà Ñ.Å. Òåçàóðóñ ïí òáîðåòè÷åñêíé è íðèëëàäííé ëèíåâèñòèéå. - Í., 1978.

Íðèêëàäíå ýçûêíçíàíèå. Ó÷åáíèê (ðåä. À.Ñ.Ãåðä). Ñíá., 1996.

Ñîéïëíà Å.Å., Èíííáíéí È.Ñ., Çàäíðóëüêí þ.À. Íöfáëåìû ïíèñáíéý êííüþòåðííé ëèíáâèñòèéè á âèäå ïòíëíâè äéý ííðòàëà çíàíéé // – I.: 2008. – N.482-487.

Þçûêîcâíàíèå. Èíôîðìàöèííí-ííèñêâûé òåçàóðóñ ÈÍÈÍ ÐÀÍ. – ï., 2007.

The Oxford handbook of computational linguistics // Mitkov Ruslan (ed.). N.Y.: Oxford university press, 2003.