Òåçàóðóñ ïî êîðïóñíîé ëèíãâèñòèêå

Àâòîð Âèêòîð Ïàâëîâè÷ Çàõàðîâ
02.09.2010 ã.
Ïîñëåäíåå îáíîâëåíèå 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 íàñòîÿùåå âðåìÿ â îáëàñòè êîðïóñíîé ëèíãâèñòèêè îòñóòñòâóåò ÷åòêàÿ ñèñòåìà íàó÷íîé òåðìèíîëîãèè. Êîðïóñíàÿ ëèíãâèñ
ïîñòðîåíèÿ è èñïîëüçîâàíèÿ ëèíãâèñòè÷åñêèõ êîðïóñîâ ñ èñïîëüçîâàíèåì êîìïüþòåðíûõ òåõíîëîãèé. Ïîä íàçâàíèåì ëèíãâèñòè÷åñêèé
âèäå è ïðåäíàçíà÷åííûé äëÿ ðåøåíèÿ êîíêðåòíûõ ëèíãâèñòè÷åñêèõ çàäà÷. Êîðïóñíàÿ ëèíãâèñòèêà íàõîäèòñÿ íà ïåðåñå÷åíèè çàä
òåðìèíîëîãèè. Âî-ïåðâûõ, ýòî åñòåñòâåííî, ó÷èòûâàÿ åå íåäàâíåå ïðîèñõîæäåíèå. Âî-âòîðûõ, êîðïóñíàÿ ëèíãâèñòèêà êàê îòäåëüíà
ñêëàäûâàòüñÿ â íåäðàõ àíãëèéñêîãî ÿçûêà. Ïîýòîìó, ðàçðàáàòûâàÿ òåçàóðóñ ïî êîðïóñíîé ëèíãâèñòèêå, åñòåñòâåííî çàäóìàòüñ
èõ àíãëèéñêèå ýêâèâàëåíòû. Â íàñòîÿùåå âðåìÿ òàêèõ òåçàóðóñîâ íå ñóùåñòâóåò.

Êðóã ëåêñèêè, îòíîñÿùåéñÿ ê ñôåðå êîðïóñíîé ëèíãâèñòèêè è ïîäëåæàùåé ôîðìàëèçàöèè, î÷åð÷åí èìåþùèìèñÿ ýíöèêëîïåäè÷åñêè
ñìåæíûì íàïðàâëåíèÿì. Ïîýòîìó ï ðè ðàçðàáîòêå òåçàóðóñà ìû èñõîäèì èç íàëè÷èÿ ðÿäà ïîñîáèé, êîòîðûå ìîãóò áûòü èñïîëüçîâàíû, à
Ëèíãâèñòè÷åñêèé 2005, ßçûêîçíàíèå 2007 è äð.].

Ïðè ïîñòðîåíèè òåçàóðóñîâ èñïîëüçóþòñÿ ðàçëè÷íûå ïîäõîäû. Îäèí èç ñïîñîáîâ èçó÷åíèÿ òåðìèíîëîãèè ïîäðàçóìåâàåò ïðåäâàðèòåëü
ðàáîòå ïðîèçâîäèòñÿ îïèñàíèå òåðìèíîëîãèè êîðïóñíîé ëèíãâèñòèêè èìåííî íà îñíîâàíèè ëîãèêî-ïîíÿòèéíîãî àíàëèçà ñ èñïîëüçîâàíèåì à
íåîáõîäèìî ñîñòàâèòü êîðïóñ òåêñòîâ ïî êîðïóñíîé ëèíãâèñòèêå è ñîçäàâàòü ñëîâíèê òåçàóðóñà íåïîñðåäñòâåííî íà æèâîì òåêñòîâ
íà ðóññêîì è àíãëèéñêîì ÿçûêàõ, êîòîðûé ïåðèîäè÷åñêè ïîïîëíÿåòñÿ íîâûìè äîêóìåíòàìè.

Ðàçðàáîòàí ïîäõîä ê ïîñòðîåíèþ òåçàóðóñîâ ïðåäìåòíûõ îáëàñòåé, ïðåäïîëàãàþùèé âûäåëåíèå êëþ÷åâûõ ïîíÿòèé èç ìíîæåñòâà èç
äåñêðèïòîðàìè, èçâëå÷åííûìè èç ó÷åáíûõ ïîñîáèé ïî êîðïóñíîé ëèíãâèñòèêå, â ÷àñòíîñòè [Çàõàðîâ 2005]. Íà âòîðîì ýòàïå ïðîâîäèëà
òåêñòû (ïðåæäå âñåãî, ïóáëèêàöèè êîíôåðåíöèé ïî êîðïóñíîé ëèíãâèñòèêå: «Êîðïóñíàÿ ëèíãâèñòèêà è ëèíãâèñòè÷åñêèå áàçû ä
– 2004, 2006, 2008» è äð.). Íà áàçå ýòîãî êîðïóñà ñîçäàåòñÿ ïåðâàÿ âåðñèÿ òåçàóðóñà ïî êîðïóñíîé ëèíãâèñòèêå. ßäðîì ý
õàðàêòåðèçóþùèõ îáëàñòü «êîðïóñíàÿ ëèíãâèñòèêà» è îòëè÷àþùèõñÿ âûñîêîé èíôîðìàòèâíîñòüþ è ÷àñòîòíîñòüþ. Äëÿ ïîñòðîåíèÿ
ëåêñèêè, ó÷èòûâàþùèé ëîãèêî-ïîíÿòèéíûå ñõåìû ïðåäìåòíîé îáëàñòè, ïðåäñòàâëåííûå â ýêñïåðòíûõ îïèñàíèÿõ. Äëÿ àâòîìàòèçèðîâà
êîëëîêàöèé, îñíîâûâàþùèåñÿ íà èçâåñòíûõ ìåðàõ àññîöèàöèè (MI, T-score, Log-Likelihood).

Òåçàóðóñ áóäåò áàçèðîâàòüñÿ íà îíòîëîãèè êîðïóñíîé ëèíãâèñòèêè, ðàçðàáàòûâàåìîé íà êàôåäðå ìàòåìàòè÷åñêîé ëèíãâèñòèêè
êîìïüþòåðíîé ëèíãâèñòèêè, ðàçðàáîòàííàÿ â Íîâîñèáèðñêå â ðàìêàõ ïðîåêòà ïî ñîçäàíèþ Ïîðòàëà çíàíèé ïî êîìïüþòåðíîé ëèíãâèñòèêå

Ðàçðàáàòûâàåìûé òåçàóðóñ äîëæåí îáåñïå÷èòü ñèñòåìàòèçèðîâàííîå ïðåäñòàâëåíèå òåðìèíîëîãèè â îáëàñòè êîðïóñíîé ëèíãâèñò
êîðïóñíîé ëèíãâèñòèêå è îáåñïå÷èò óäîáíûé äîñòóï ê íèì. Äâóÿçû÷íîñòü òåçàóðóñà ïîçâîëèò îòå÷åñòâåííûì ó÷åíûì è ñïåöèàëèñòà
ïåðåâîäå ñòàòåé.

Òåçàóðóñ áóäåò ïðåäñòàâëÿòü ñîáîé ýëåêòðîííóþ áàçó äàííûõ, îáåñïå÷èâàþùóþ ïîëüçîâàòåëÿì äîñòóï ê íåìó. Êðîìå òîãî, áóäåò ñ
ìíîãîÿçû÷íûõ (ISO 5964-1985, ÃÎÑÒ Ð 7.24-2007) òåçàóðóñîâ.

Ëèòåðàòóðà

Áàðàíîâ À.Í. Ââåäåíèå â ïðèêëàäíóþ ëèíãâèñòèêó. Ñåðèÿ "Íîâûé ëèíãâèñòè÷åñêèé ó÷åáíèê". Ì.: Ýäèòîðèàë ÓÐÐÑ. 2001.

Áîðîâèêîâà Î.È., Çàãîðóëüêî Þ.À., Çàãîðóëüêî Ã.Á., Êîíîíåíêî È.Ñ., Ñîêîëîâà Å.Ã. Ðàçðàáîòêà ïîðòàëà çíàíèé ïî êîìïüþòåðíîé ëèíãâ
(ã.Äóáíà, Ðîññèÿ). –Ì.: ËÅÍÀÍÄ, 2008. –Ò.3. –Ñ.380-388.

Âèíîãðàäîâà Í.Â., Ìèòðîôàíîâà Î.À. Ôîðìàëüíàÿ îíòîëîãèÿ êàê èíñòðóìåíò ñèñòåìàòèçàöèè äàííûõ â ðóññêîÿçû÷íîì êîðïóñå òåêñòîâ ï
ÑÏá.: ÑÏáÃÓ, 2008. – C. 113-121.

Äåìüÿíêîâ Â.Ç. Àíãëî-ðóññêèå òåðìèíû ïî ïðèêëàäíîé ëèíãâèñòèêå è àâòîìàòè÷åñêîé ïåðåðàáîòêå òåêñòà. Âûï. 2. Ìåòîäû àíàëèçà

Çàõàðîâ Â.Ï. Êîðïóñíàÿ ëèíãâèñòèêà: Ó÷åáíî-ìåòîäè÷åñêîå ïîñîáèå. – ÑÏá.: ÑÏáÃÓ, 2005. – 48 ñ.

Ëèíãâèñòè÷åñêèé ýíöèêëîïåäè÷åñêèé ñëîâàðü. Ì.: Ñîâ. Ýíöèêëîïåäèÿ, 1990.

Ìèòðîôàíîâà Î.À., Çàõàðîâ Â.Ï. Àâòîìàòèçèðîâàííûé àíàëèç òåðìèíîëîãèè â ðóññêîÿçû÷íîì êîðïóñå òåêñòîâ ïî êîðïóñíîé ëèíãâèñòèêå
«Äèàëîã–2009». Ì.: 2009.

Íèêèòèíà Ñ.Å. Òåçàóðóñ ïî òåîðåòè÷åñêîé è ïðèêëàäíîé ëèíãâèñòèêå. - Ì., 1978.

Ïðèêëàäíîå ÿçûêîçíàíèå. Ó÷åáíèê (ðåä. À.Ñ.Ãåðä). ÑÏá., 1996.

Ñîêîëîâà Å.Ã., Êîíîíåíêî È.Ñ., Çàãîðóëüêî Þ.À. Ïðîáëåìû îïèñàíèÿ êîìïüþòåðíîé ëèíãâèñòèêè â âèäå îíòîëîãèè äëÿ ïîðòàëà çíàíèé //
– Ì.: 2008. – Ñ.482-487.

ßçûêîçíàíèå. Èíôîðìàöèîííî-ïîèñêîâûé òåçàóðóñ ÈÍÈÎÍ ÐÀÍ. – Ì., 2007.

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