Ïîñòðîåíèå âåðîÿòíîñòíîé ìîäåëè äëÿ ñëîãîâîé ñòðóêòóðû òåêñòà.

Àâòîð Ìèõàèë Àëåêñàíäðîâè÷ Øàðîíîâ
11.07.2010 ã.
Ïîñëåäíÿÿ îáíîâëåíèå 08.10.2010 ã.

The objects of study are the statistics of Russian text, which obtained from informants for the syllabic markup and programmatically for characters of the text. The text is submitted as sequences that consist from random variables - characters of the text. The main order of research is the influence statistical characteristics of chars to the statistical characteristics of composite events presented as sequences of letters, which in linguistics can be defined through the notion of syllable. The main idea of that study is identifying influence chars statistic's, for statistic's of composite events, such as syllables. Established regularities proposed to use in applied problems of speech's modeling, in particular, in defining the boundaries between syllables.

Ïðè ðàáîòå â ïîãðàíè÷íûõ ñ ëèíãâèñòè÷åñêèìè íàïðàâëåíèÿìè îáëàñòÿõ çíàíèé ÷àñòî ïðèõîäèòñÿ ñòàëêèâàòüñÿ ñ ïðîáëåìàìè òîëê
áóêâà. Ñóùåñòâóþò ðàçëè÷íûå óòâåðæäåíèÿ î êàæäîì èç ýòèõ ïîíÿòèé, ÷òî óêàçûâàåò íà íàëè÷èå ìîæåñòâåííé è âàæíé ïðîáëå
íàèáîëåå ïîäõîäÿùèå, ïî íàøåìó ìíåíèþ, äëÿ áíàëèçà ñëîãîâîé îðãàíèçàöèè òåêñòà.

Ðàññìîòðèì ýòè îñíîâíûå ïîðåäåëåíèÿ.

Ñëîâî (â ãðàôè÷åñêîì àñïåêòå) – ïîñëåäîâàòåëüíîñòü çíàêîâ, îãðàíè÷åííàÿ ïðîáåëàìè;

Ìîðôåìà – ýëåìåíòàðíàÿ äâóñòîðîííÿÿ ÷àñòü (êîìïîíåíò) ñëîâà, «â êîòîðîé ça ïîðåäåëåííé ôîíåòè÷åñêé ôîðìîé (îçíà÷àþùèì) çà

Îïðåäåëåíèå äëÿ ïîíÿòèÿ ñëîãà åù¸ áîëåå íåîäíîçíà÷íî. Ïî ñâåäåíèÿì [Âåñòíèê 2010, 1/1: 100] âèäíî, ÷òî òàêîå ÿâëåíèå êàê ñëîãà, íåñò
àñïåêòàõ: àêóñòè÷åñêèõ, àðòèêóëÿöèîííûõ, ôóíêöèîíàëüíûõ. Äëÿ óäîáñòâà â äàëüíåéøåì èñïîëüçóåòñÿ ñëåäóþùàÿ èíòåðïðåòàö

«Ñëîã – ýòî îáúåäèíåíèå ñîãëàñíûõ è ãëàñíûõ â ãðóïïó. Ñëîã êàê ôîíîëîãè÷åñêàÿ åäèíèöà îïðåäåëÿåòñÿ êàê ìèíèìàëüíàÿ ã
ñòðóêòóðîé èçó÷àåìîãî ÿçûêà» [Âåñòíèê 2010, 1/1: 101].

Áóêâà – ýëåìåíòàðíàÿ (äàëåå íåäåëèìàÿ) åäèíèöà ãðàôè÷åñêîé ñèñòåìû ÿçûêà, ïðåäñòàâëÿþùàÿ ñîáîé êëàññ ãðàôè÷åñê

Ôîíåìà – (ïî àíàëîãèè ñ áóêâîé) ýòî òàêàÿ åäèíèöà çâóêîâîé ñèñòåìû ÿçûêà, êîòîðàÿ ïðåäñòàâëÿåò ñîáîé êëàññ ôîíåòè÷åñ
ëèíãâèñòîâ, ñïåöèôè÷åí è, âîîáùå ãîâîðÿ, íåîäíîçíà÷åí. Åñëè èìåòü â âèäó «÷èñòîãî» ëèíãâèñòà — â êà÷åñòâå «êðàéíåãî
ëèíãâèñòà — ýòî òåêñòû. Îäíàêî äàæå ïðèìåíèòåëüíî ê äàííîé ñèòóàöèè íåëüçÿ ñêàçàòü, ÷òî òåêñòû è ÿâëÿþòñÿ ïîäëèííûì
å. ñèñòåìó àáñòðàêòíûõ, íáùèõ äëÿ âñåõ òåêñòîâ ÿëåìåíòîâ è çàêîíîìåðíîñòåé èõ ôóíêöèîíèðîâàíèÿ â òåêñòàõ. Òàê ìîæíî ïîèñàòü
÷òîáû ïîèñàòü ÿçûê ïî äàííûì òåêñòà» [Ùåðáà 1974].

Êàê óïîìèíàåòñÿ â [Ôîíîëîãè÷åñêèå 1983], «…â ðóññêîì ÿçûêå ÷ëåíåíèå ñëîâà íà ìèíèìàëüíûå çíà÷èìûå åäèíèöû (ìîðôåìû) è íà
ñëîâà ‘îòûñêàëàñü’: ‘îò-èñê-à-ë-à-ñü’; ‘/à-ôûc-êà-ëàñ/’. Ãðàíèöû ìîðôåì è ñëîãîâ çäåñü
âíóòðè ðàçíûõ ñëîãîâ, ñîåäèíÿÿñü ñ ÷àñòÿìè ñîñåäíèõ ìîðôåì, ò.å. èìååò ìåñòî ñëîãîâîå ïåðåðàçëîæåíèå ìîðôåì».

Âàæíûì ýòàïîì â êîíñòðóèðîâàíèè ñèñòåì îáðàáîòêè ðå÷è ÿâëÿåòñÿ ýòàï ïðîâêòèðîâàíèÿ ìåõàíèçìà ïîðåäåëåíèÿ ãðàíèö ìåæäó ñëîã
íàïðàâëåíèé íàóê, èçó÷àþùèõ ðå÷åâûå ìåõàíèçìû. Ñ êëàññèôèêàöèåé ñòðàòåãèé ÷èñòàâëåíèé ìîæíî îçíàêîìèòüñÿ â

Äëÿ ðåøåíèÿ çàäà÷è ïîñòðîåíèÿ èíòåðôåéñà ñèñòåìû àäåêâàòíîãî îáùåíèÿ (äàëåå ÑÀÎ) [Ïèñüìåííå 2009: 71] òàêæå íåîáõîäèìî ïðà
òàêèõ ãðàíèö ïîìîãàåò â ïîðåäåëåíèè ñìûñëîâîé íàãðóçêè ñëîâà.

Íà ïðàêòèêå ãðàíèöû ìåæäó ñëîãàìè íåîäíîçíà÷íû è âíóòðè îäíîãî è òîãî æå ñëîâà ìîãóò âûãëÿäåòü, íàïðèìåð, ñëåäóþùèì îáðàçîì: &
íûé’, ‘ïîñ-ìàò-ðè-âàë’. Ïðàâàÿ ÷àñòü äàííîãî ïðèìåðà äåìîíñòðèðóåò ãðàíèöû ñëîãîâîé ðàçìåòêè áåç ó÷¸òà ìîð
ñëîâ.

Ïåðåõîäÿ ê êîíñòðóèðîâàíèþ èíòåðôåéñà ÑÀÎ, íåîáõîäèìî îïðåäåëèòü, êàêèì îáðàçîì íåôîðìàëèçîâàííàÿ èíôîðìàöèÿ äîëæíà ïðåäñòàâ
óæå èìåþùåéñÿ â ïàìÿòè è ò.ï.

Êðîìå òîãî, ïðè ðàçðàáîòêå àëãîðèòìà ðàáîòû èíòåðôåéñà ÑÀÎ íåîáõîäèìî ó÷åñòü, ÷òî ðå÷åâûå ñîîáùåíèÿ äîëæíû ïðåäñòàâëÿòüñÿ

Â äàííé ðàáîòå ìåõàíèçì ðàçãðàíè÷åíèÿ ñëîãîâ â òåêñòå èññëåäîâàëñÿ ìåòîäîì ýêñïåðòíûõ îöåíîê. Â êà÷åñòâå ìàòåðèàëà äëÿ èñ
ïîâåñòè îáú¸ìì áîëåå 3000 ñèìâîëîâ íà ñëîãè, ðóêîâîäñòâóÿñü ëè÷íûì ïîòòîì.

Áûëè ïîäñ÷èòàíû ÷àñòîòíûå õàðàêòåðèñòèêè ñëîãîâ, âûäåëåííûõ ýêñïåðòàìè. Ñâîäíàÿ òàáëèöà âûáîðî÷íûõ ñâåäåíèé ïî äâóì ýêñï

Ìîæíî îòìåòèòü, ÷òî â íåêîòîðûõ ñëó÷àÿõ âåðîÿòíîñòè äåëåíèÿ ñëîâà íà ñëîãè îêàçàëèñü ðàçíûì. Íàïðèìåð: îäèí èç ýêñïåðòîâ ñìîã ð
äåëåíèè ñëîâà íà ñëîãè [ñì. Ïðèëîæåíèå 1].

Àíàëèç âñåõ òèïîâ ñëîãîâ, ïîêàçàë íàëè÷èå áîëüøîãî êîëè÷åñòâà ðàçëè÷íûõ òèïîâ ñëîãîâ ñ îäèíàêîâûìè âåðîÿòíîñòÿìè ïîÿâëåíèÿ [ñ
âåðîÿòíîñòåé èõ ïîÿâëåíèÿ. Â äàëüíåéøåì ïðè ïîñòðîåíèè ñòàòèñòè÷åñêèõ ìîäåëåé òåêñòîâ èñïîëüçîâàëèñü óñëîâíûå âåðîÿòíîñòè

Ñëåäóåò îòìåòèòü, ÷òî ïðèâåä¸ííûå â òàáëèöå ïðèëîæåíèÿ 1 âàðèàíòû ñëîãîâ õàðàêòåðèçóþòñÿ, â ïåðâóþ î÷åðåäü, óäîáñòâîì çâ
òèïîâ ñëîãîâ â ðå÷åâûå öåïî÷êè [Âåñòíèê 2010, 1/1: 102].

Òàêèì îáðàçîì, ìîäåëèðîâàíèå ÿçûêîâûõ ïðîöåññîâ ïîçâîëÿåò âûÿâèòü îñíîâíûå òðåáîâàíèÿ ê èíòåðôåéñó ÑÀÎ, îðèåíòèðóÿñü íà ñëîã

Íàðÿäó ñ áóêâåííûì àëôàâèòîì, â ðàáîòå ïðåäëàãàåòñÿ èñïîëüçîâàòü àëôàâèò, ñîñòîÿùèé èç ñëîãîâûõ ñòðóêòóð, îòîæäåñòâëÿÿ
ðàññìàòðèâàòüñÿ êàê îñíîâà äëÿ ïîñòðîåíèÿ ÿçûêîâîé ìîäåëè, ó÷èòûâàþùåé êðèòåðèè èíòóèòèâíîãî îáðàçîâàíèÿ ðå÷åâûõ ñîîáùåí

Ïðè ðàçãðàíè÷åíèè ñëîãîâ öåëåñîîáðàçíî ó÷èòûâàòü ñìûñëîâûå íàãðóçêè, êîòîðûå ïåðåäàíû â ìîðôåìàõ. Ïðè ðàçðûâå ìîðôåìû ñëîãî

Èñïîëüçîâàíèå âåðîÿòíîñòíûõ õàðàêòåðèñòèê áóêâ, ïðèâåä¸ííûõ â ïðèëîæåíèè 2, äëÿ âû÷èñëåíèÿ óñëîâíîé âåðîÿòíîñòè ïîñëåäîâà
áóêâ â ñëîãå îïðåäåëåíà êîíêðåòíûìè îñîáåííîñòÿìè èçó÷àåìîãî ÿçûêà, è â ÷àñíûõ ñëó÷àÿõ, å¸ ìîæíî ðàññìàòðèâàòü êàê íåêîòîðóþ

Ëèòåðàòóðà

- Âåñòíèê 2010 - Âåñòíèê ÌÃÎÓ «Ëèíãâèñòèêà». Ì: Íóâêà, 2010;

- Êîìïüþòåðíîå 2009 – Êîðî÷êîâ À.Â. Êîìïüþòåðíîå ìîäåëèðîâàíèå ãðàôåìíî-ôîíåìíîãî ïðåîáàçîâàíèÿ â àíãëèéñêîì ÿçûêå. Ñàð

- Îáùàÿ 2001 – Êîäçàñîâ Ñ.Â. Îáùàÿ ôîíåòèêà. Ì: ðîñ. ãîñ. ãóìàíèò. óí-ò, 2001;

- Ïèñüìåííîå 2009 – Áàðàíîâ Â.À. Ïèñüìåííîå íàñëåäèå è ñîâðåìåííûå èíôîðìàöèîííûå òåõíîëîãèè. Èæââñê, 2009;

- Ôîíîëîãè÷åñêèå 1983 – Êàñåâè÷ Â.Á. Ôîíîëîãè÷åñêèå ïðîáëåìû îáùåãî è âîñòî÷íîãî ÿçûêîçíàíèÿ. Ì., 1983;

- Ùåðáà 1974 - Ùåðáà Ë. Â. ßçûêîâàÿ ñèñòåìà è ðå÷åâàÿ äåÿòåëüíîñòü. Ì.: Íàóêà, 1974.