A R A N E A : ÑÅÌÅÉÑÒÂÎ ÌÈËËÈÀÐÄÍÛÕ ÂÅÁ-ÊÎÐÏÓÑÎÂ

© Âëàäèì]ð Áåíêî (Vladimír Benko). Ñëîâàêèÿ, Áðàòèñëàâà. Èíñòèòóò
ÿçûêîçíàíèÿ èì. Ëþäîâèòà Øòóðà Ñëîâàöêîé
Àêàäåìèè íàóê, ÞÍÅÑÊÎ
êàôåäðà ìíîãîÿçû÷íîé è ìåæêóëüòóðíîé êîììóíèêàöèè Óíèâåðñèòåòà èì. ßíà
Êîìåíñêîãî â Áðàòèñëàâå

Ëåêöèè ïðåäñòàâÿò ïðîåêò
ñåìåéñòâà âåá-êîðïóñîâ Aranea äëÿ ÿçûêîâ èñïîëüçîâàííûõ è/èëè
ïðåïîäàâàííûõ â ñëîâàöêèõ óíèâåðñèòåòàõ, êîòîðûå ïðåäíàçíà÷åíû äëÿ ïðåïîäàâàíèÿ
ôèëîëîãè÷åñêèõ è òðàíñëàòîëîãè÷åñêèõ ïðåäìåòîâ, à òàêæå è äëÿ ëèíãâèñòè÷åñêèõ
èññëåäîâàíèé.

Âåá-êîðïóñ ïðåäñòàâëÿåò ñîáîé îñîáûé
âèä ëèíãâèñòè÷åñêîãî êîðïóñà, êîòîðûé ñîçäàí ïóòåì ïîñòåïåííîé çàãðóçêè òåêñòîâ
èç èíòåðíåòà ïðè ïîìîùè àâòîìàòèçèðîâàííûõ ïðîöåäóð, êîòîðûå íà ëåòó îïðåäåëÿþò
ÿçûê è êîäèðîâêó îòäåëüíûõ âåá ñòðàíèö, óäàëÿþò øàáëîíû, ýëåìåíòû íàâèãàöèè,
ññûëêè è ðåêëàìó (ò.í. boilerplate),
îñóùåñòâëÿþò òðàíñôîðìàöèþ íà òåêñò, ôèëüòðàöèþ, íîðìàëèçàöèþ è äåäóïëèêàöèþ
ïîëó÷åííûõ äîêóìåíòîâ, êîòîðûå çàòåì ìîæíî îáðàáîòàòü òðàäèöèîííûìè
èíñòðóìåíòàìè êîðïóñíîé ëèíãâèñòèêè (òîêåíèçàöèÿ, ìîðôîñèíòàêñè÷åñêàÿ è
ñèíòàêñè÷åñêàÿ àííîòàöèÿ) è âíåäðèòü â ïîèñêîâóþ êîðïóñíóþ ñèñòåìó. Ñîçäàíèå
âåá-êîðïóñà íå òîëüêî íàìíîãî äåøåâëå, íî ïðåæäå âñåãî åãî ðàçìåð ìîæåò áûòü
äàæå íà ïîðÿäîê áîëüøå òðàäèöèîííûõ êîðïóñîâ.

Ïðè ñîçäàíèè âñåõ êîðïóñîâ áûëà
ïðèìåíåíà îäèíàêîâàÿ ìåòîäîëîãèÿ è îäèíàêîâûé íàáîð ïðîãðàìíûõ èíñòðóìåíòîâ: SpiderLing, Onion, Unitok [1] è TreeTagger [2]. Â êà÷å
ïîèñêîâîé ñèñòåìû èñïîëüçóåòñÿ NoSketchEngine [3] (open-ource) èëè SketchEngine [4] (ïëàòíàÿ). Ó êîðïóñîâ íàçâàíèÿ íà «íåéò
(ëàòèíñêîì) ÿçûêå îáîçíà÷àþùåì ÿçûê è òîæå ðàçìåð êîðïóñà, íàïðèìåð AraneumAnglicumMinus,
AraneumRussicumMaius, è ò. ï. Â íàñòîÿùåå âðåìÿ ñåìåéñòâî ñîäåðæèò
18 êîðïóñîâ íà 14 ÿçûêàõ â äâóõ ðàçìåðàõ è âñå êîðïóñû äîñòóïíû â áåñïëàòíîì
ðåæèìå íà êîðïóñíîì ïîðòàëå ïðîåêòà [5].

Â îòëè÷èå îò âû÷èñëèòåëüíûõ ëèíãâèñòîâ,
êîòîðûå îáðàáàòûâàþò êîðïóñíûå äàííûå â ïàêåòíîì ðåæèìå, îñòàëüíûå ïîëüçîâàòåëè
êîðïóñà îáû÷íî çàèíòåðåñîâàíû â ïîèñêàõ êîíêðåòíûõ ïðèìåðîâ ñëîâ, ñî÷åòàíèé è
ñèíòàêñè÷åñêèõ ñòðóêòóð ðàñïîëîæåííûõ â âèäå êîíêîðäàíñîâ, ÷àñòîòíûõ ñïèñêîâ è
ïðîôèëåé íà ýêðàíå. Ó÷èòûâàÿ ðàçìåðû ñîâðåìåííûõ êîðïóñîâ ñòàíåò ÿñíûì, ÷òî
ýôôåêòèâíîñòü è óäîáíîñòü ïîèñêîâîé ÿâëÿåòñÿ î÷åíü âàæíûì ôàêòîðîì äëÿ âñÿêîé
ðàáîòû ñ êîðïóñîì.

Ñåìèíàð ïîñâÿùåí ïðàêòèêå ðàáîòû ñ
ñåìåéñòâîì êîðïóñíûõ ïîèñêîâûõ ñèñòåì NoSketch
Engine [6]è Sketch Engine [7],
ïðèíàäëåæàùèì ê ñàìûì ëó÷øèì â ìèðå èíñòðóìåíòàì äëÿ ðàáîòû ñî «ñâåðõáîëüøèìè»
êîðïóñàìè (ðàçìåðîì â äåñÿòêè ìèëëèàðäîâ òîêåíîâ). Îáå ñèñòåìû áûëè ñîçäàíû â Ëàáîðàòîðèè îáðàáîòêè åñòåñòâåííîãî ÿçûêà
Ôàêóëüòåòà èíôîðìàòèêè Óíèâåðñèòåòà èì. Ìàñàðèêà â Áðíî, ïðè÷åì ôóíêöèè
áåñïëàòíîé (open-source) ñèñòåìû NoSketch
Engine ÿâëÿþòñÿ ïîäìíîæåñòâîì ôóíêöèé Sketch Engine è âêëþ÷àþò â ïîëíîì
îáúåìå ðàáîòó ñî ñïèñêàìè ñëîâ (Word List)
è êîíêîðäàíñåðîì, ò. å. ïîèñê ïî ñëîâîôîðìå, ëåììå, ñî÷åòàíèè è ìîðôîñèíòàêñè÷åñêîé
ìåòêå â ðàçíûõ êîìáèíàöèÿõ íà ÿçûêå CQL
(Corpus Query Language), è òîæå

âû÷èñëåíèå êîëëîêàöèé íà áàçå ñòàòèñòè÷åñêèõ ìåð ñîº÷åòàåìîñòè (T-score, MI, MI3, log likelihood, min. sensitivity è logDice). Ñèſòåìà ðàáîòàåò â ðåæèìå ñåðâåð/êëèåíò, ãäå íà ñåðâåðå õðàíÿòſ âſå äàííûå è îſóùåſòâëÿþſſ ïîèſêîâûå ſïåðàöèè è ïîëüçîâàòåëü ðàáîòàåò ſ êëèåſòſ ÷åðåç âåá-èſòåðôåſſ ïðè ſſſîè ſòàíäàðòſſîãſ áðàóçåðà.

Ïëàòſàÿ ſèſſòåſà Sketch Engine ſſſäåðæèò êðſîå âſåõ ôóſêöèé NoSketch Engine òðè ſóùåſòâåſſûõ ðàſſøèðåſèÿ – êſîëſſêſ ſðſſîôèëè (ſêâòſ÷è) ſſſòſſîåſſûå íà áàçå ſſëüçſſâàòåëüſêſé ſêâòſ÷-ãðàſſàòèſè, äèſòðèáóòèâſûé òåçàóðſſ è ôóſêöèþ ſðàâſåſèÿ ſêâòſ÷åé äëÿ äâóõ ëåſſſ÷åſêſõ åäèſíë. Âſſå ýòſ ôóſêöèſ ðàáſòàþſ ſ äàſſûìſ âû÷èſëåſſûìſ çàðàſſå, ſòſ äåëàåſò ſſſòſàſ ſ÷åſü áſſòðſé è óäſáſſé. Ñſſſòåſà ſðåäſſſòàâëÿåſſÿ â âſäå ſſåðâſſà (ſſäſſſſêſ) íà ſſåðåſðàſ êſſſſäſſſ Lexical Computing, íà êſòſſðſſſ ſſſðſſóſſû áſſëſå ſ÷ſ íà 80 ſçſſſàſ, âſſëþ÷àſ 15-ſſëſſſàðäſſé êſſðſſóſ ðóſſſſêſãſ ſçſſêà.

Êſſſêſðäàſſ ſſſðſàòà KWIC äëÿ ſſſſâſſſðſû «ſſâſſſſáſðſſſé»

Ïſàâſſòſſðſſſå
êſſſſêſàòſ ſſſſâſſſðſû «ſſâſſſſáſðſſſé»

Äſſòſðſáóòſâſſé òåçàâſ äëÿ ëåſſû
«ſſâſſſſáſðſſſé»

×àſſòſſàÿ äſſòſðſáóſſÿ ëåſſû
«ſſâſſſſáſðſſſé» ïſ TLD

Ïſſſãſàſſà
êóſſà

«A r a n e a : Ñåſåéſòâſ ſſſſſàðäſſûſ
âåá-êſſðſóſſâ»

(10 ÷àſſâ ſåêſſſſſſûſ çàſÿòſé è ſſàêſſſſóſâ,
16-20 ſſÿáðÿ 2015 ã.)

Ëåêſſÿ 1. Âſåäſåſſå. Ëſſſãâſſòſ÷åſſſé êſðſóſ êàê ſſòſ÷ſſê ſſſôſðſàſſſ ſ ÿçſſêå

Îſſſſâſſûå ſſſÿſſÿ: âſſäſ ýſåêſſðſſſûſ êſſſſåêſſé òåſſòſâ

Èſſſðſÿ ſſſçäàſſÿ êſðſóſſâ, ãåſſåðàſſè, ſðſſſåſåſÿ

Ïðîåêòû íàöèîíàëüíûõ êîðïóñîâ

Êîðïóñíàÿ ëèíãâèñòèêà êàê ìåòîä èññëåäîâàíèÿ âåòõà
ÿçûêîçíàíèÿ

Ïðèìåíåíèå ìåòîäîâ êîðïóñíîé ëèíãâèñòèêè â ñèíõðîííîì è
äèàõðîííîì èññëåäîâàíèè ÿçûêà

Ïîíÿòèå âåá-êîðïóñà, îñíîâåííòè è îòëè÷èÿ îò
òðàäèöèîííûõ êîðïóñîâ

Ëåêöèÿ 2 Aranea – Ñåìåéñòâî ìèëëèàðäíûõ
âåá-êîðïóñîâ

Îñíîâíûå ðåøåíèÿ ïðîåêòà Aranea: ÿçûêè, ðàçìåðû è âàðèàíòû, íàçâàíèÿ

Èíñòðóìåíòû äëÿ îáðàáîòêè: êðàóëèíã, ïîðåäåëåíèå ÿçûêà, óäàëåíèå
øàáëîíîâ, äåïóëèêàöèÿ, òîêåíèçàöèÿ, ìîðôîñèíòàêñè÷åñêàÿ ðàçìåòêà,óíèôèêàöèÿ
òåãñåòîâ

Ïóáëèêàöèÿ è èñïîëüçîâàíèå êîðïóñîâ: êîðïóñíûå ìåíåäæåðû

Ôîðìàò êîðïóñîâ Aranea [8]:
àòðèáóòû è ñòðóêòóðû

Ïðàêòèêóì:

A. Ðàáîòà
ñ êîðïóñíîé ïîèñêîâîé ñèñòåìîé (No)SketchEngine

Ïðàêòèêóì

1. Ïîèñê ïî ñëîâîôîðìå, ëåììå è
ñëîâîñî÷åòàíèþ

Ðåæèì èçîáðàæåíèÿ, íàñòðîéêè

Ôèëüòðû, êîíòåêñò

×àñòîòíûå äèñòðèáóöèè

Òåãñåòû (tagsets), ïîèñê ïî òåãàì

2. ßçûê çàïðîñîâ CQL

Ðåãóëÿðíûå âûðàæåíèÿ

Ïðèìåíåíèå CQL:
ïîiñê ñèíòàêñè÷åñêèõ ñòðóêòóð

Êîëëîêàöèè, ìåðû àññîöèàòèâíîñòè

Âû÷èñëåíèå êîëëîêàöèîííûõ êàíäèäàòîâ

B. Ðàáîòà
ñ êîðïóñíîé ïîiñêîâîé ñèñòåìîé SketchEngine

Ïðàêòèêóì 1: Êîðïóñíûé ìåíåäæåð Sketch Enginå (https://www.sketchengine.co.uk/)

Êîëëîêàöèîííûå ïðîôèëè (ñêåò÷è)

Ñêåò÷-ãðàììàòèêè: ñèíòàêñè÷åñêèé è êîëëîêàöèîííûé ïîäõîä

Ñêeò÷-ãðàììàòèêè äëÿ êîðïóñîâ Aranea

Ñðàâíåíèå ñêåò÷åé (Sketch-diff)
è äèñòðèáóòèâíûé òåçàâð

Äâóÿçû÷íûå ñêåò÷è

Ïðàêòèêóì
2: Ðåñóðñû è èíòóìåíòû ñàéòà Sketch
Engine

Êîðïóñû ñåìåéñòâà TenTen

Ïàðàëëåëüíûå êîðïóñû

Èíñòðóìåíòû äëÿ ñîçäàíèÿ ïîëüçîâàòåëüñêèõ êîðïóñîâ: Corpus
Architect è WebBootCaT

Ýêñòðàêöèÿ òåðìèíîëîãèè

Â ðàìêàõ
ïðàêòèêóìîâ ñòóäåíòû îçíàêîìÿòñÿ ñ ðàáîòîé ñ îáèèì ñèñòåìàìè è ïîëó÷àò
íåîãðàíè÷åííûé äîñòóï ê êîðïóñàì íà êîðïóñíîì ïîðòàëå ïðîåêòà Aranea (http://ucts.uniba.sk/) è âðåìåííûé 3-õ
ìåñÿ÷íûé áåñïëàòíûé àêêàóíò äëÿ Sketch
Engine.

Webography

https://savba.academia.edu/VladimirBenko

http://ucts.uniba.sk/aranea_about/

[1]
http://corpus.tools/

[2]
http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[3]
http://nlp.fi.muni.cz/trac/noske

[4]
https://www.sketchengine.co.uk/

[5]
http://ucts.uniba.sk/aranea_about/

[6]http://nlp.fi.muni.cz/trac/noske

[7]https://www.sketchengine.co.uk/

[8] Ñàéò ïðîåêòà Aranea: http://ucts.uniba.sk/aranea_about/

Êîðïóñíûå ïîðòàëû: Aranea: http://ucts.uniba.sk/
; http://ella.juls.savba.sk/