Araneum Bohemicum Maximum: Veľký český webový korpus

Datum konání: 28. 4. 2017

Místo konání: Filozofická fakulta, budova G, učebna G13

Název přednášky: Araneum Bohemicum Maximum: Veľký český webový korpus

Přednášející: Ing. Vladimír Benko, PhD.

Počet účastníků: 18

Zpracoval: Radek Kopečný

Dne 28. 4. 2017 v počítačové učebně G13 FF MU jsme se zúčastnili již třetí přednášky v rámci jarního přednáškového cyklu externích odborníků. Nyní byl hostem Ing. Vladimír Benko, PhD., který nám nejdříve představil korpus Araneum Bohemicum Maximum, a v druhé části přednášky jsme si mohli sami vyzkoušet práci s korpusy ve Sketch Engine.

Ing. Vladimír Benko, PhD.

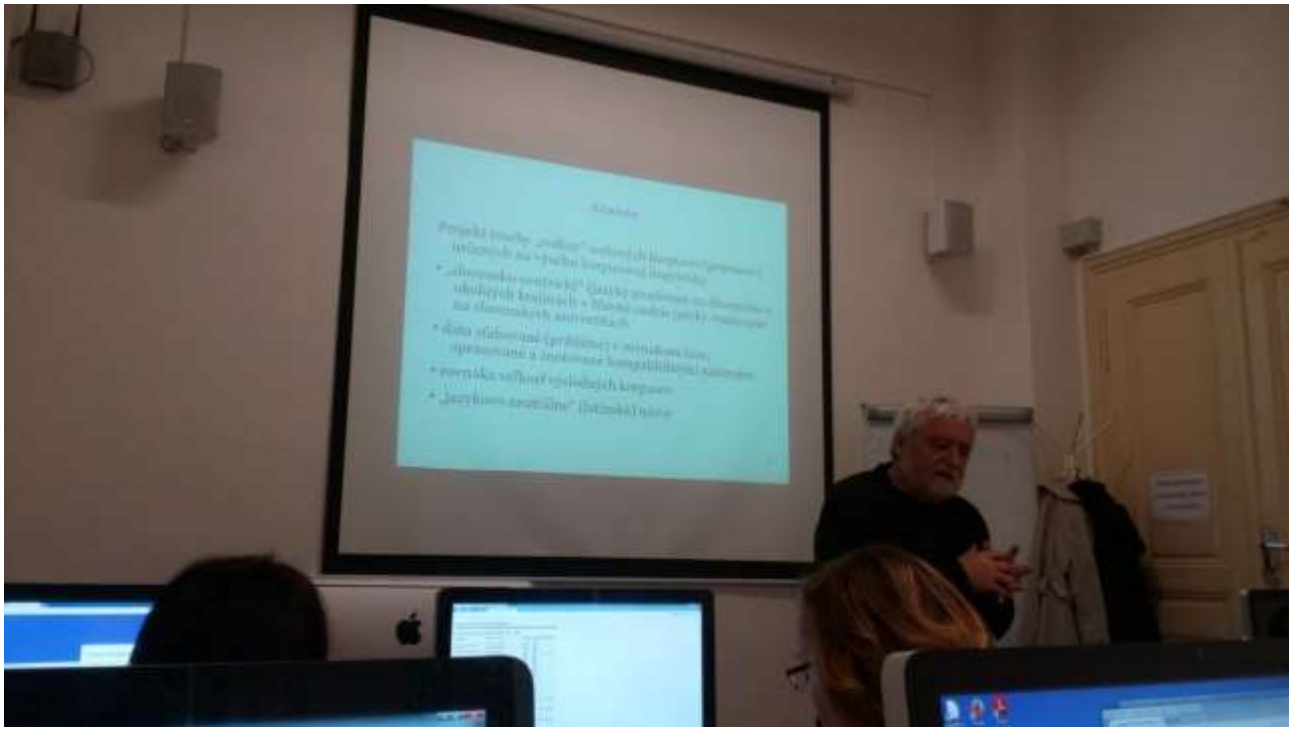
Ing. Vladimír Benko, PhD., působí jako odborný pracovník v Jazykovědném ústavu Ľudovíta Štúra na Slovenské akademii věd v Bratislavě. Mezi jeho projekty patří *Slovník súčasného slovenského jazyka – 5. etapa (Koncipovanie a redigovanie slovníkových hesiel a s tým spojený lexikologicko-lexikografický výskum)* a korpus Aranea, na kterém, jak jsme se dozvěděli, pracuje sám. Mezi jeho oblasti výzkumu patří také morfosyntaktická anotace, WSG (Word Sketch Grammar) a počítačová lexikologie a lexikografie.

První část přednášky

Ing. Vladimír Benko, PhD., nám nejdříve představil historii korpusů. Vůbec první korpus

na světě – Brown Corpus – vznikal v 60. letech minulého století a obsahoval milion slov, zachycuje americkou angličtinu. Dalším korpusem byl britský LOB s podobnou strukturou jako Brown Corpus, jen se zaměřením na britskou angličtinu. Z českých nám byly představeny například korpusy řady SYN nebo TenTen.

Poté se Ing. Vladimír Benko, PhD., zaměřil na představení toho, jak vzniká webový korpus, jenž obsahuje články z internetu. Nejprve dojde ke stažení dat z internetu a provede se konverze těchto dat na text (odstranění značek HTML a skriptů). Následně dochází k odstranění „orámování“ textu (boilerplate) a identifikaci a normalizaci kódování. Poté je identifikován samotný jazyk a cizojazyčné texty jsou vyfiltrovány. Následnou „deduplikací“ dojde k odstranění duplicitních textů a také dochází k filtraci defektních textů (absence diakritiky, nesprávně interpretované kódování, …).



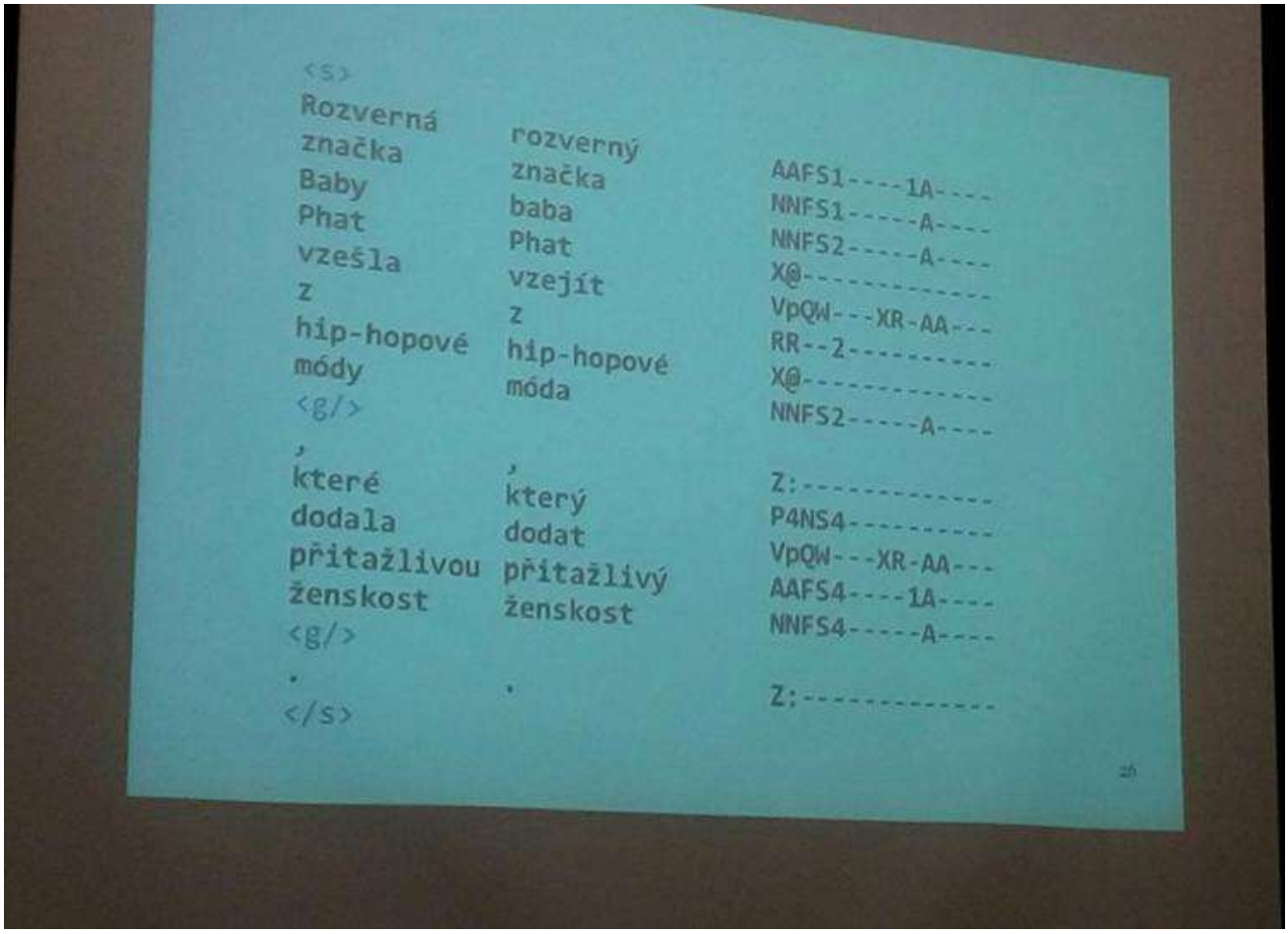
Po uvedení webových korpusů obecně jsme se dostali k samotnému projektu Aranea, jenž je projektem tvorby „rodiny“ webových korpusů primárně určených k výuce počítačové lingvistiky. Aranea je tzv. „slovensko-centrický“, zaměřený hlavně na jazyky používané na Slovensku a v okolních zemích + cizí jazyky, jež jsou vyučované na slovenských univerzitách. Stahovaná data pro jednotlivé jazyky probíhají zhruba ve stejném období, jsou zpracována a anotována kompatibilními nástroji a je zde snaha o stejnou velikost výsledných korpusů. Zajímavostí je používání latinských jazykově neutrálních názvů.

Araneum Bohemicum Maximum patří k největším korpusům v projektu. Stahování dat

probíhalo v pěti obdobích – 2013, 2x 2015 a 2x 2016 pomocí nástroje SpiderLing.

Program Unitok následně provádí tokenizaci (rozčlenění textu složeného z písmen, mezer a interpunkčních znamének na tokeny). Morfologickou analýzu textů provádí program MorphoDita. Závěrem se texty segmentují na věty, filtrují, provádí se konverze tagsetu Evy Hajičové na tagset AUT (Araneum Universal Tagset) a k tomu používá Ing. Vladimír Benko, PhD., vlastní nástroje.

Jak taková analýza probíhá, jsme si ukázali na větě: „*Rozverná značka Baby Phat vzešla z hip-hopové módy, které dodala přitažlivou ženskost*.“



Druhá část – workshop

Druhá část přednášky měla charakter workshopu. Vyzkoušeli jsme si sami, jak pracovat s korpusy Aranea a poté i Sketch Engine. Korpsuy Aranea http://aranea.juls.savba.sk/ jsou veřejně přístupné pod účtem hosta s jistými omezeními, ale dostali jsme možnost se registrovat pro plný přístup a Ing. Vladimír Benko, PhD., nám naši registraci potvrdil.

Káva Araneum Bohemicum III Minus (Czech, 17.04) 120 M

Collocation candidates

Page 1 Go Next >

		Cooccurrence count	Candidate count	T-score	MI	logDice
P | N	**šálek**	194	890	13.925	11.913	9.694
P | N	**zrnkový**	77	94	8.774	13.823	8.519
P | N	**pití**	78	3,413	8.810	8.659	7.970
P | N	**ranní**	75	4,145	8.633	8.322	7.813
P | N	**pražený**	45	283	6.706	11.458	7.705
P | N	**pít**	72	5,736	8.447	7.795	7.558
P | N	**zelený**	112	14,256	10.507	7.119	7.447
P | N	**mletý**	39	951	6.236	9.503	7.368
P | N	**plantážní**	30	36	5.477	13.848	7.172
P | N	**milovník**	47	4,185	6.821	7.634	7.134
P | N	**instantní**	30	343	5.474	10.596	7.108
P | N	**odpolední**	36	3,040	5.971	7.711	6.908
P | N	**cibetkový**	23	31	4.795	13.680	6.789
P | N	**ledový**	30	2,761	5.449	7.587	6.687
P | N	**příprava**	84	20,455	9.039	6.183	6.659
P | N	**italský**	37	6,379	6.023	6.681	6.525
P | N	**výborný**	50	11,726	6.977	6.237	6.468
P | N	**lahodný**	21	1,042	4.570	8.478	6.458
P | N	**čerstvý**	38	7,396	6.097	6.506	6.456
P | N	**uvařit**	22	1,790	4.669	7.764	6.394
P | N	**silný**	77	24,604	8.616	5.791	6.329

Na lemmatu káva jsme si ukázali jeden ze způsobů využití korpusu – hledání kolokací.

milovat/nenávidět Araneum Bohemicum III Maius (Czech, 17

milovat 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 nenávidět

X/Y , X/Y	13,878	1,781	0.13	0.12
zachovávat	59	0	7.0	--
žít	138	0	6.9	--
meditovat	50	0	6.8	--
odpouštět	39	0	6.4	--
toužit	41	0	6.3	--
ctít	31	0	6.1	--
obdivovat	33	0	6.1	--
dávat	63	0	6.1	--
věřit	70	0	6.0	--
cítit	61	0	5.9	--
respektovat	31	0	5.9	--
poznat	53	0	5.9	--
radovat	27	0	5.8	--
přijímat	32	0	5.8	--
trpět	33	0	5.7	--
dělat	235	15	6.5	2.8
znát	91	10	6.0	3.2
dokázat	85	9	6.2	3.5
milovat	1,222	142	10.5	8.2
snášet	58	20	6.8	7.2
nenávidět	142	108	8.2	10.0
závidět	0	12	--	6.9
odsuzovat	0	12	--	7.0
proklínat	0	10	--	7.3
pohrdat	0	12	--	7.4

Y X	48,229	6,980	0.45	0.48
bezmezně	240	0	7.3	--
svobodně	231	0	7.1	--
nadevše	186	0	7.0	--
heidi	155	0	6.7	--
tajně	108	0	6.0	--
doopravdy	111	0	5.8	--
vroucně	82	0	5.8	--
bezpodmínečně	69	0	5.4	--
smírně	84	0	5.3	--
prostě	696	28	5.9	1.4
opravdu	796	43	5.5	1.4
skutečně	387	22	5.5	1.6
bůh	685	59	6.8	3.5
tolik	513	72	6.4	3.9
hluboce	158	16	6.2	4.2
přestat	179	33	5.3	3.4
doslova	187	41	5.7	4.2
strašně	148	33	5.7	4.4
vášnivě	109	9	6.1	5.0
navzájem	124	38	5.7	5.1
upřímně	93	104	5.5	7.0
duše	40	212	3.5	6.6
svorně	0	18	--	6.0
spam	0	45	--	7.0
bytostně	0	56	--	7.6

Následně jsme si ukazovali a vyzkoušeli nástroj Sketch Engine (dále SE), jenž vyvíjí společnost Lexical Computing Ltd. ve spolupráci s Centrem zpracování přirozeného jazyka Fakulty informatiky Masarykovy univerzity, a proto k němu máme bezplatný přístup. SE slouží k vytváření

korpusů a vyhledávání v nich, má přístup k mnoha korpusům ve více než 80 jazycích a například Aranea mezi ně též patří. Oproti běžným nástrojům má rozšířené možnosti, například Thesaurus (spojení slov na základě jejich podobnosti ve významu – SE ještě vyobrazí jejich frekvenci výskytu v konkrétním korpusu). Právě díky mnoha jazykům je SE velmi užitečný nástroj pro překladatele, neboť umožňuje vyhledávat kolokace paralelně, dělí je dle četnosti použití, dle pozic (za slovem, nebo před slovem) a také dle slovních druhů. Zajímavé je také vyhledávání dvojic slov a možnost podívat se, se kterými slovy se více spojuje první z dvojice, se kterými to druhé a která slova mají společná (viz obrázek).

Na závěr byl prostor pro dotazy na samotný projekt Araneum i korpusy obecně. Přednáška byla velmi zajímavá a díky širokému využití korpusů si přišli na své i studenti jiného oboru než počítačové lingvistiky.