Někdo rozumí počítačům. Počítače nerozumí nikomu. Naučíme je to?

Corpus-Based Knowledge

Otázka: **Kdo ví víc?**

Odpověď: **Bůh ví 4,2578537320× častěji než čert.**

Vytváříme obrovské soubory textů (jazykové korpusy), ze kterých se lingvisté o jazyku leccos dozvídají: která slova se používají často, málo často, často spolu, která slova jsou nová a co znamenají. Vývoj software, který umožňuje indexovat miliardy slov a pomocí regulárních výrazů v nich rychle vyhledávat, je informatická výzva.

vědět

czTenTen12 [Majka] frekvence = 5931659 (1091.0 v miliónu)

has_subj	440718	-8.1	post_prep	265921	-4.5	coord	91437	-1.7
fakt	19690	9.3	o	165643	8.1	znát	6088	7.24
bůh	16942	8.83	od	8189	4.96	tušit	830	6.81
moc	29095	8.68	z	18031	4.65	tápat	332	6.67
čert	3979	8.03	díky	643	4.36	chtít	2209	6.41
dávno	3580	7.8	podle	1554	3.79	neznat	904	6.41
rada	9698	7.29	prostřednictvím	231	3.7	vědět	4488	6.23
člověk	31112	7.23	kvůli	416	3.67	umět	1462	6.1
málo	3868	7.15	za	4609	3.53	rozumět	936	6.08
prd	1625	6.83	s	9095	3.5	chápat	1053	5.88

Centrum zpracování přirozeného jazyka
Fakulta informatiky
Masarykova univerzita