ЭКСПЕРИМЕНТ С СОВМЕСТИМОЙ СКЕТЧ-ГРАММАТИКОЙ

Владимир Бенко

vladob@juls.savba.sk

Словацкая академия наук
Институт языкознания им. Людовита Штура
Братислава

«Корпусная лингвистика»
Санкт-Петербург, июнь 2013

Word Sketches

Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour.

[Kilgarriff et al., Euralex 2004]

Word Sketches

Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour.

[Kilgarriff et al., Euralex 2004]

Word Sketches

Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour.

[Kilgarriff et al., Euralex 2004]

Word sketch … «эскиз слова» … коллокационный профиль

Коллокационные профили

- **Корпус текстов с морфологической (или морфосинтактической) разметкой**
- **Программа Word Sketch Engine**
- **Набор правил (скетч-грамматика) созданных с учетом нужд пользователей корпуса**

Скетч-грамматика

- **Основана на языке CQL (Corpus Query Language)**

```
[tag="A.*"] []{0,3}   [tag="N.*"]
```

Скетч-грамматика

- **Основана на языке CQL (Corpus Query Language)**

```
2:[tag="A.*"] []{0,3} 1:[tag="N.*"]
```

Скетч-грамматика

- **Основана на языке CQL (Corpus Query Language)**

```
=a_modifier
2:[tag="A.*"] []{0,3} 1:[tag="N.*"]
```

Скетч-грамматика

- **Ключевое слово:** **словарь**

a_modifier	260656	0.3
толковый	43173	11.95
энциклопедический	35926	11.8
англо-русский	4412	9.05
немецко-русский	3250	8.62
этимологический	3213	8.61
русско-немецкий	3159	8.6
русско-английский	3147	8.59
биографический	3347	8.48
Англо-русский	2411	8.22
орфографический	2271	8.07
терминологический	2184	8.01
краткий	4906	7.72
философский	3550	7.58
толково-фразеологический	1074	7.07

Скетч-грамматика

Наш подход

- **Названия правил не означают синтаксические отношения, а коллокационные**
- **Позиция ключевого слова и коллоката показана явно**
- **Часть речи ключевого слова не определяется**

Скетч-грамматика

Наш подход

Av X/X Av

Vb X/X Vb

Aj X, X Aj

Sb X, X Sb

Z X, X Z

Pp X, X Pp

X/Y Cj X/Y

Pp X Y, X Pp Y

Pp Y X, Y Pp X

Aj(X), Vb(X)…

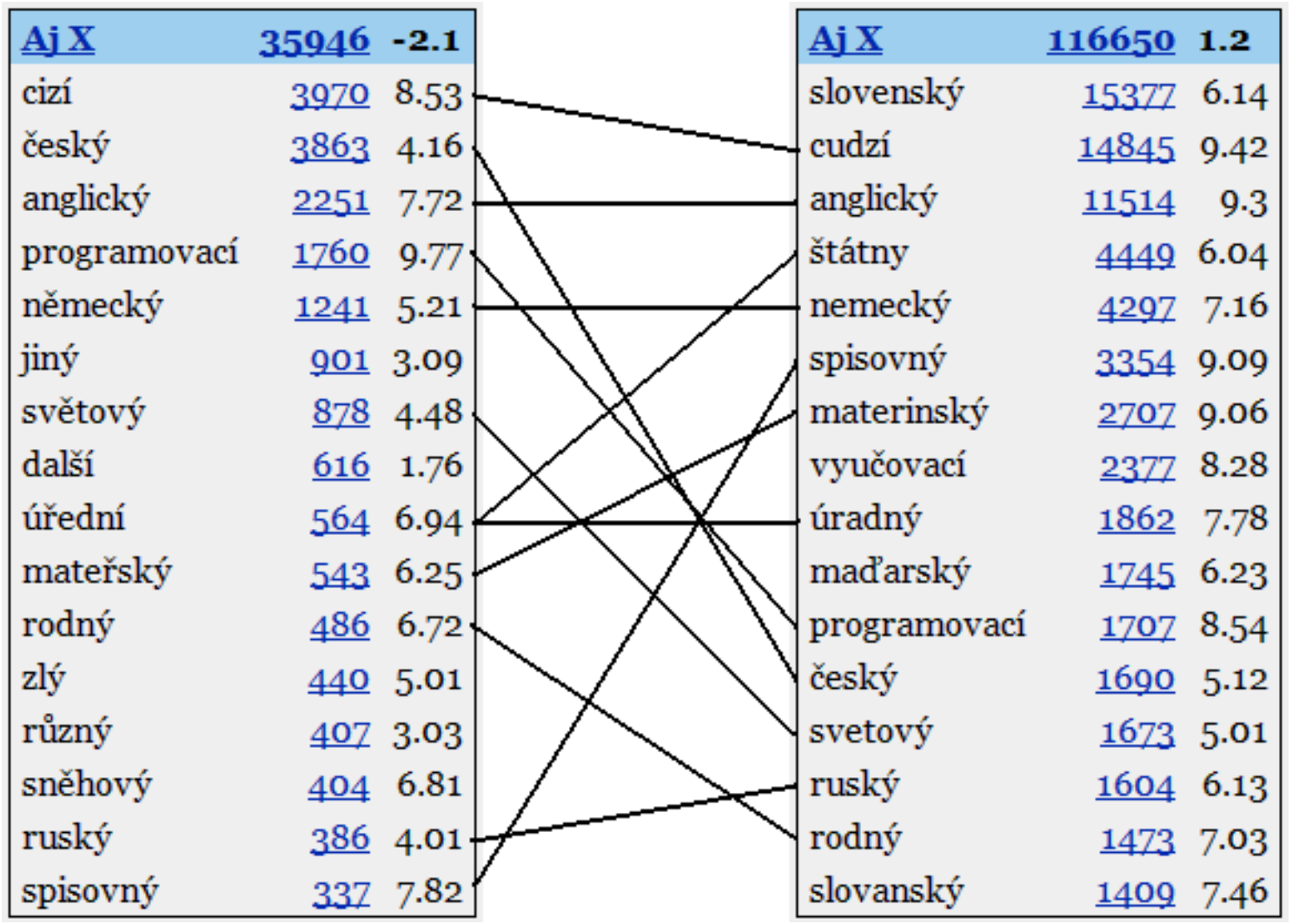
nom(X), pl(x)…

Применение скетчей в Институте языкознания им. ЛШ

- **Словарь современного словацкого языка (8-томный, 2 тома уже опубликованы)**
- **Коллокационный словарь (1-ый том – Коллокации с существительными именами – подготовлен к печати)**
- **Чешско-словацкий словарь (2-томный, работы пока в начале)**

Параллельные скетчи

jazyk (cs) / jazyk (sk)

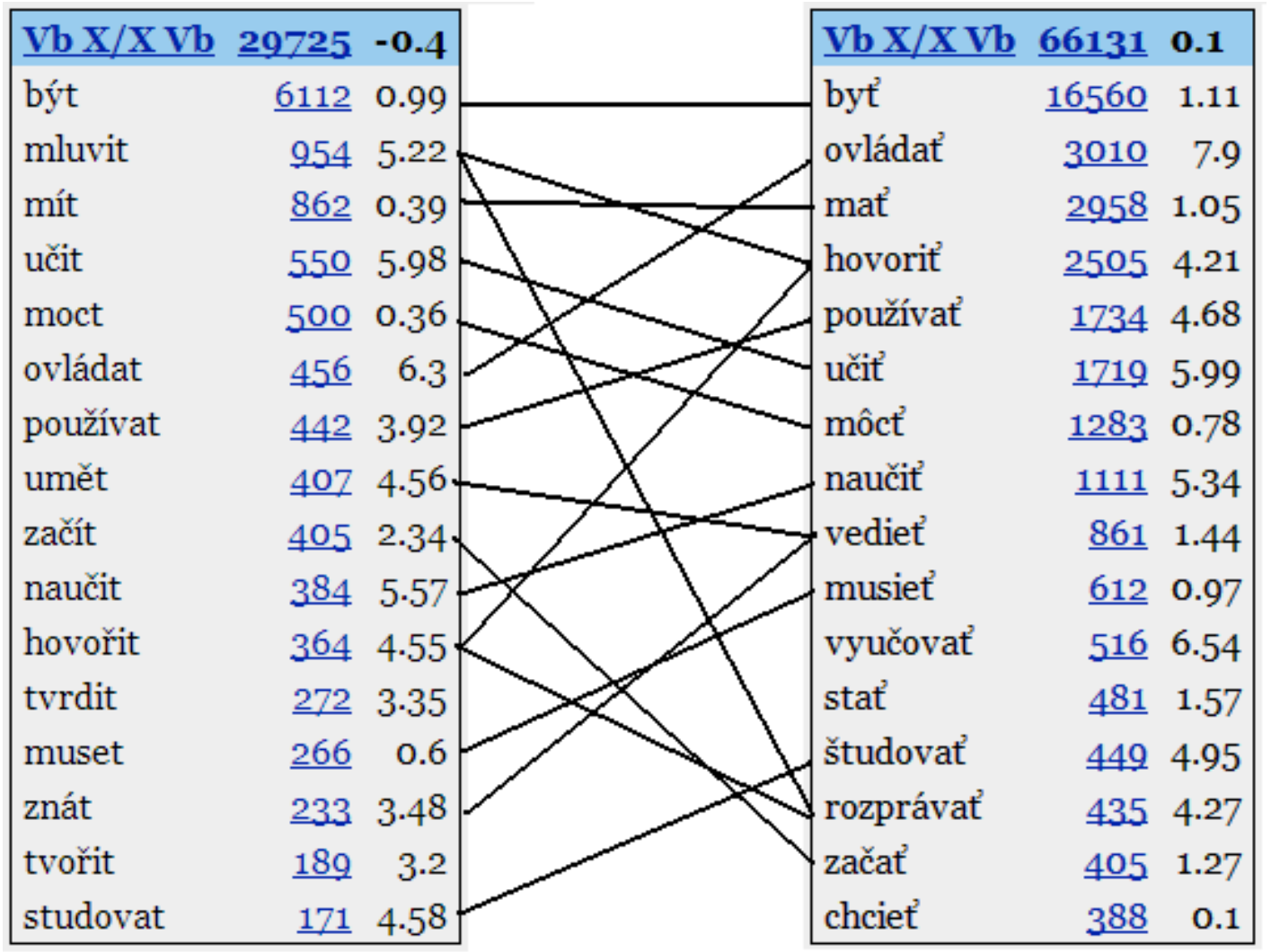


Aj X	35946	-2.1
cizí	3970	8.53
český	3863	4.16
anglický	2251	7.72
programovací	1760	9.77
německý	1241	5.21
jiný	901	3.09
světový	878	4.48
další	616	1.76
úřední	564	6.94
mateřský	543	6.25
rodný	486	6.72
zlý	440	5.01
různý	407	3.03
sněhový	404	6.81
ruský	386	4.01
spisovný	337	7.82

Aj X	116650	1.2
slovenský	15377	6.14
cudzí	14845	9.42
anglický	11514	9.3
štátny	4449	6.04
nemecký	4297	7.16
spisovný	3354	9.09
materinský	2707	9.06
vyučovací	2377	8.28
úradný	1862	7.78
maďarský	1745	6.23
programovací	1707	8.54
český	1690	5.12
svetový	1673	5.01
ruský	1604	6.13
rodný	1473	7.03
slovanský	1409	7.46

Параллельные скетчи

jazyk (cs) / jazyk (sk)

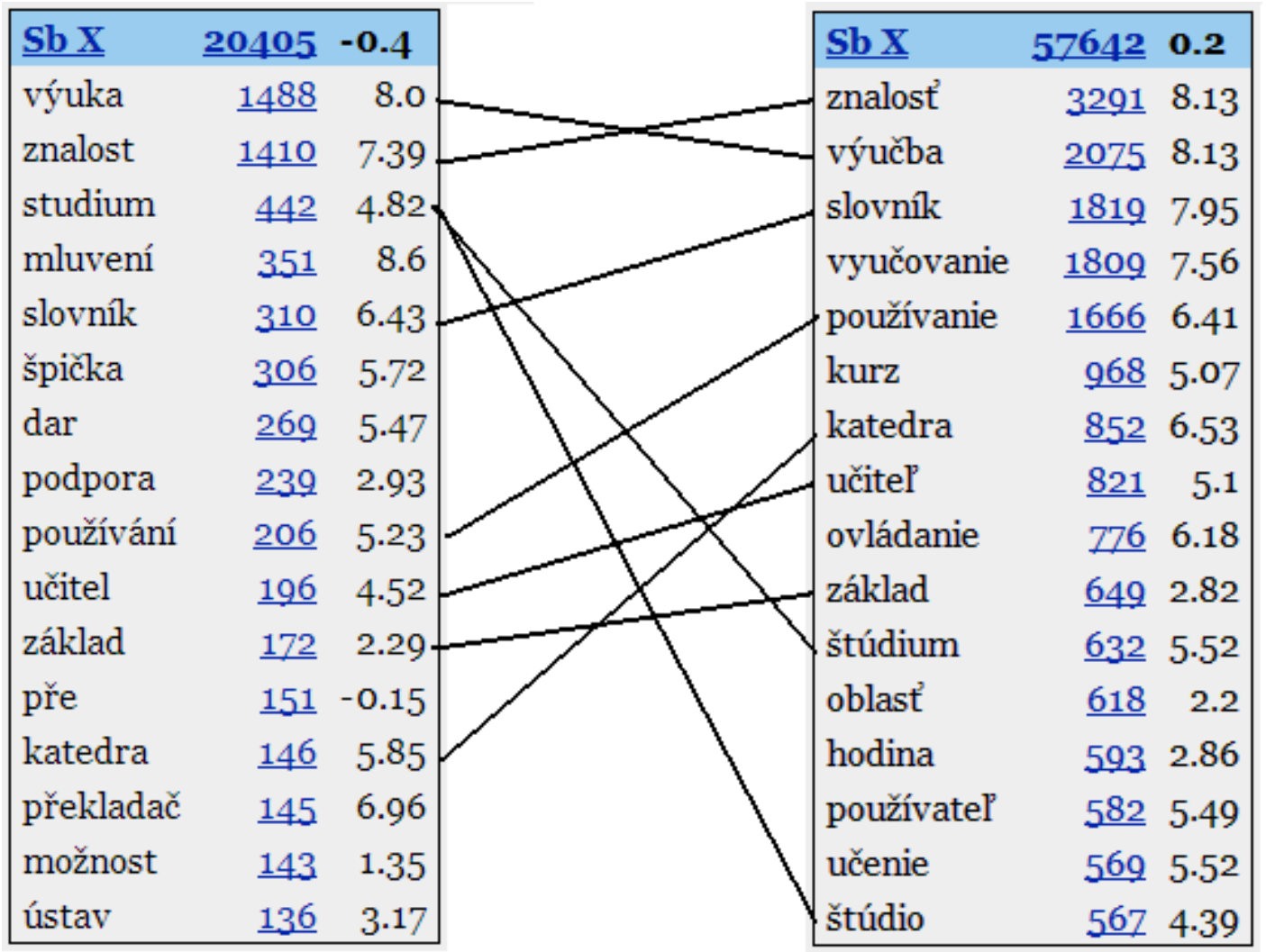


Vb X/X Vb	29725	-0.4
být	6112	0.99
mluvit	954	5.22
mít	862	0.39
učit	550	5.98
moct	500	0.36
ovládat	456	6.3
používat	442	3.92
umět	407	4.56
začít	405	2.34
naučit	384	5.57
hovořit	364	4.55
tvrdit	272	3.35
muset	266	0.6
znát	233	3.48
tvořit	189	3.2
studovat	171	4.58

Vb X/X Vb	66131	0.1
byť	16560	1.11
ovládať	3010	7.9
mať	2958	1.05
hovoriť	2505	4.21
používať	1734	4.68
učiť	1719	5.99
môcť	1283	0.78
naučiť	1111	5.34
vedieť	861	1.44
musieť	612	0.97
vyučovať	516	6.54
stať	481	1.57
študovať	449	4.95
rozprávať	435	4.27
začať	405	1.27
chcieť	388	0.1

Параллельные скетчи

jazyk (cs) / jazyk (sk)



Sb X	20405	-0.4
výuka	1488	8.0
znalost	1410	7.39
studium	442	4.82
mluvení	351	8.6
slovník	310	6.43
špička	306	5.72
dar	269	5.47
podpora	239	2.93
používání	206	5.23
učitel	196	4.52
základ	172	2.29
pře	151	-0.15
katedra	146	5.85
překladač	145	6.96
možnost	143	1.35
ústav	136	3.17

Sb X	57642	0.2
znalosť	3291	8.13
výučba	2075	8.13
slovník	1819	7.95
vyučovanie	1809	7.56
používanie	1666	6.41
kurz	968	5.07
katedra	852	6.53
učiteľ	821	5.1
ovládanie	776	6.18
základ	649	2.82
štúdium	632	5.52
oblasť	618	2.2
hodina	593	2.86
používateľ	582	5.49
učenie	569	5.52
štúdio	567	4.39

Русский веб-корпус

- **Веб-краулер SpiderLing, оптимизированный для скачивания текстов (модуль для определения языка документа, удаление шаблонов)**
- **Морфологическая разметка: TreeTagger & MULTEXT-East тегсет**
- **Дедупликация документов методом фингерпринтов (дубликаты удалены)**

Русский веб-корпус

- **Сегментация предложений**
- **Дедупликация предложений (дубликаты обозначены)**
- **Унификация тегов для пунктуации**
- **Перенос словацкой скетч-грамматики на русский тегсет MULTEXT-East**
- **Обработка с помощью WSE**

Русский веб-корпус

- 7,9 Гбайт сырого текста
- 18,9 Гбайт текста в вертикальном формате (после морфологиоческой разметки)
- 604 тысяч документов
- 41,3 млн предложений
- 699 млн токенов (словоформ, включая пунктуацию)

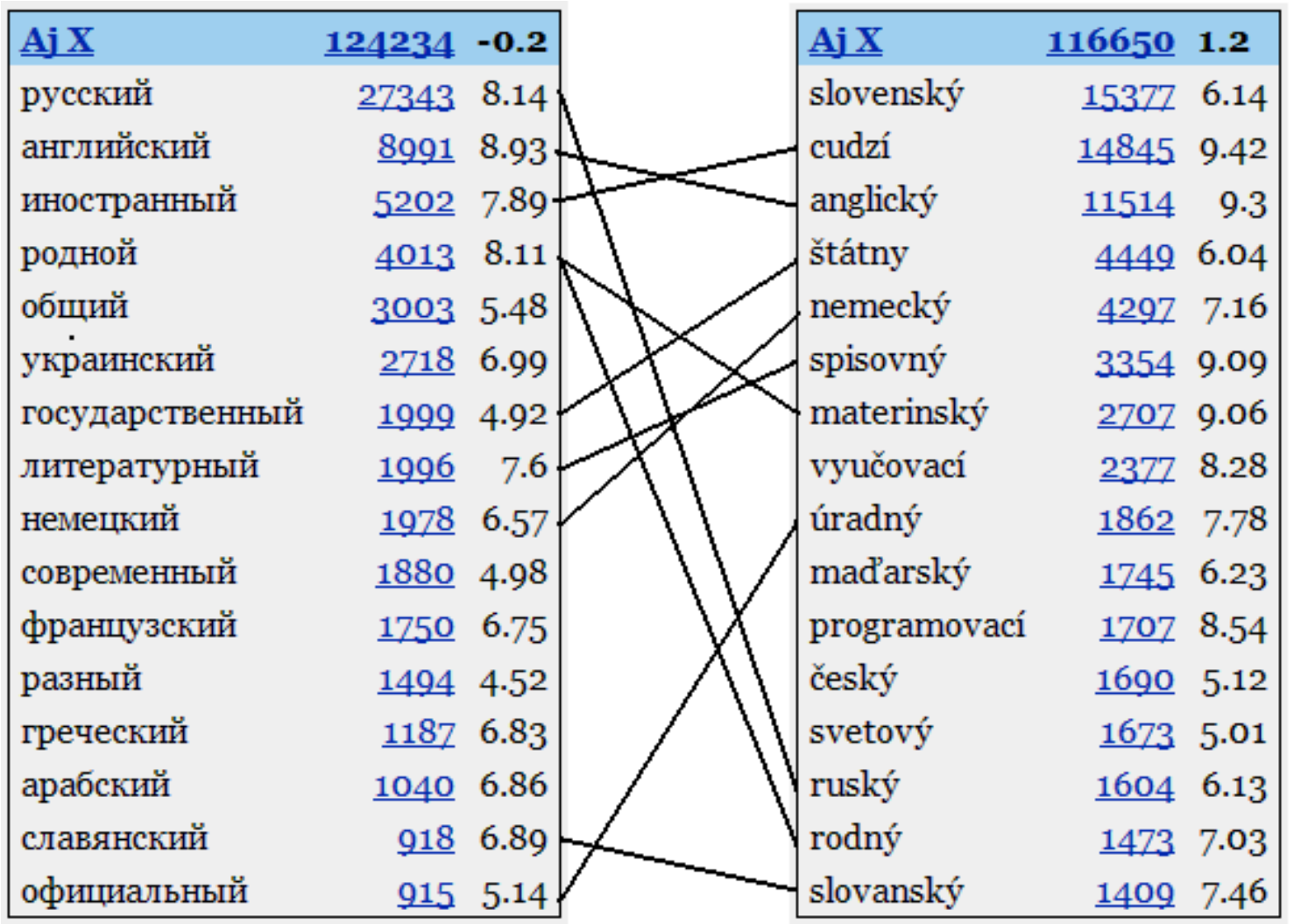
Русский веб-корпус

Дедупликация

- **Удалено 15,9 тысяч документов (2,7%)**
- **Выявлено дублетами и обозначено 11,1 млн предложений (27,6%) содержащих 155 млн токенов (22,1%)**

Параллельные скетчи

язык (ru) / jazyk (sk)

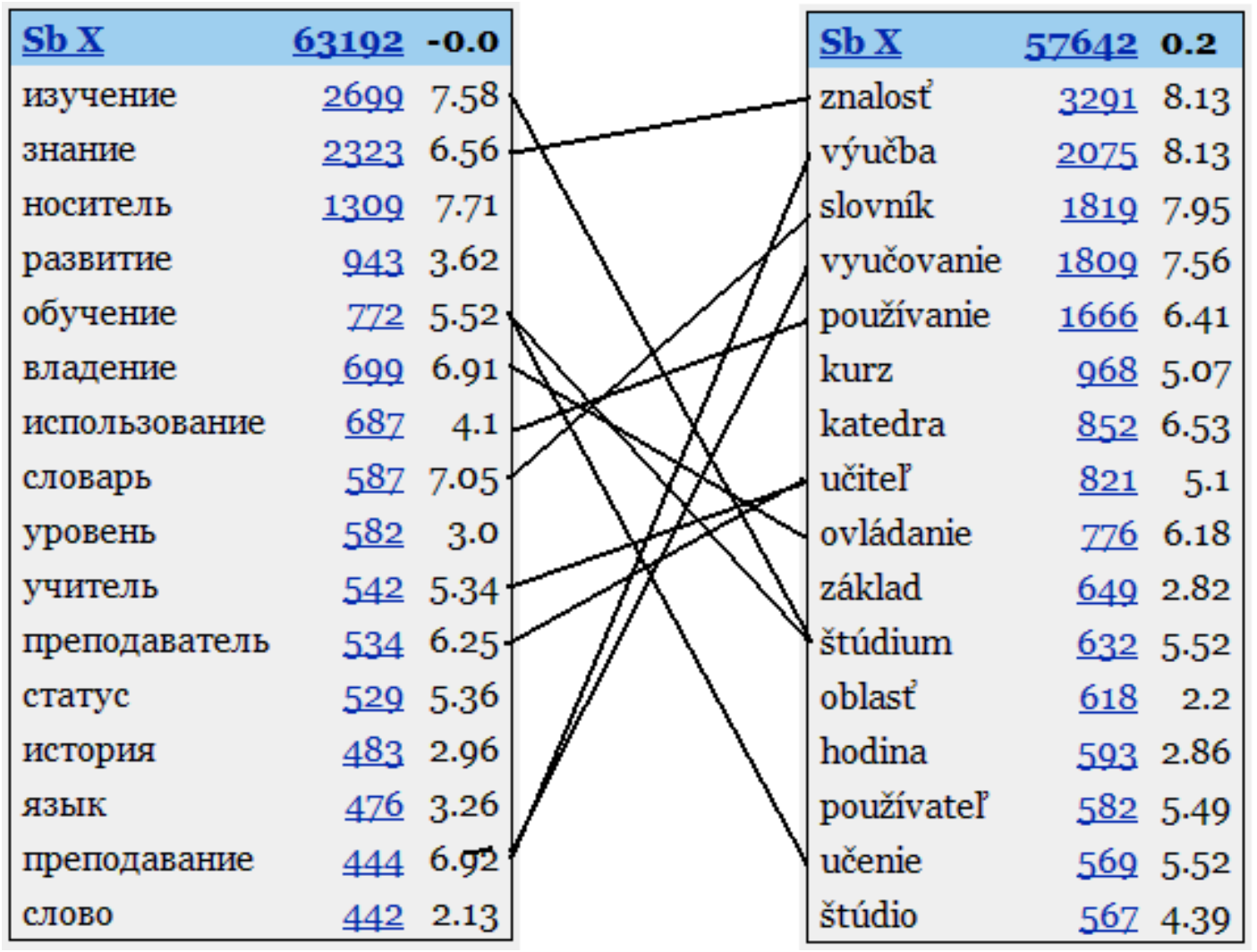


Aj X	124234	-0.2
русский	27343	8.14
английский	8991	8.93
иностранный	5202	7.89
родной	4013	8.11
общий	3003	5.48
украинский	2718	6.99
государственный	1999	4.92
литературный	1996	7.6
немецкий	1978	6.57
современный	1880	4.98
французский	1750	6.75
разный	1494	4.52
греческий	1187	6.83
арабский	1040	6.86
славянский	918	6.89
официальный	915	5.14

Aj X	116650	1.2
slovenský	15377	6.14
cudzí	14845	9.42
anglický	11514	9.3
štátny	4449	6.04
nemecký	4297	7.16
spisovný	3354	9.09
materinský	2707	9.06
vyučovací	2377	8.28
úradný	1862	7.78
maďarský	1745	6.23
programovací	1707	8.54
český	1690	5.12
svetový	1673	5.01
ruský	1604	6.13
rodný	1473	7.03
slovanský	1409	7.46

Параллельные скетчи

язык (ru) / jazyk (sk)

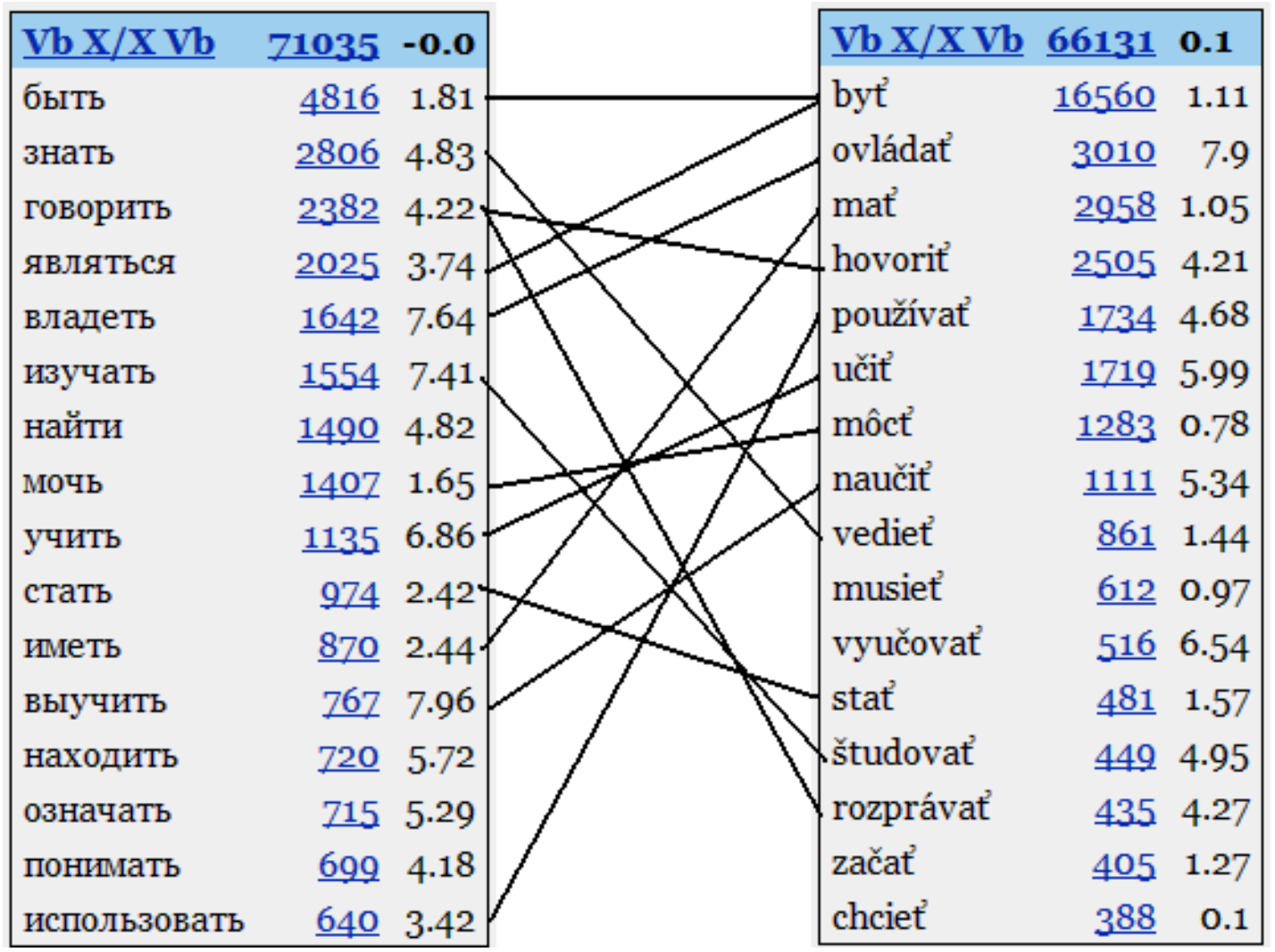


Sb X	63192	-0.0
изучение	2699	7.58
знание	2323	6.56
носитель	1309	7.71
развитие	943	3.62
обучение	772	5.52
владение	699	6.91
использование	687	4.1
словарь	587	7.05
уровень	582	3.0
учитель	542	5.34
преподаватель	534	6.25
статус	529	5.36
история	483	2.96
язык	476	3.26
преподавание	444	6.92
слово	442	2.13

Sb X	57642	0.2
znalosť	3291	8.13
výučba	2075	8.13
slovník	1819	7.95
vyučovanie	1809	7.56
používanie	1666	6.41
kurz	968	5.07
katedra	852	6.53
učiteľ	821	5.1
ovládanie	776	6.18
základ	649	2.82
štúdium	632	5.52
oblasť	618	2.2
hodina	593	2.86
používateľ	582	5.49
učenie	569	5.52
štúdio	567	4.39

Параллельные скетчи

язык (ru) / jazyk (sk)

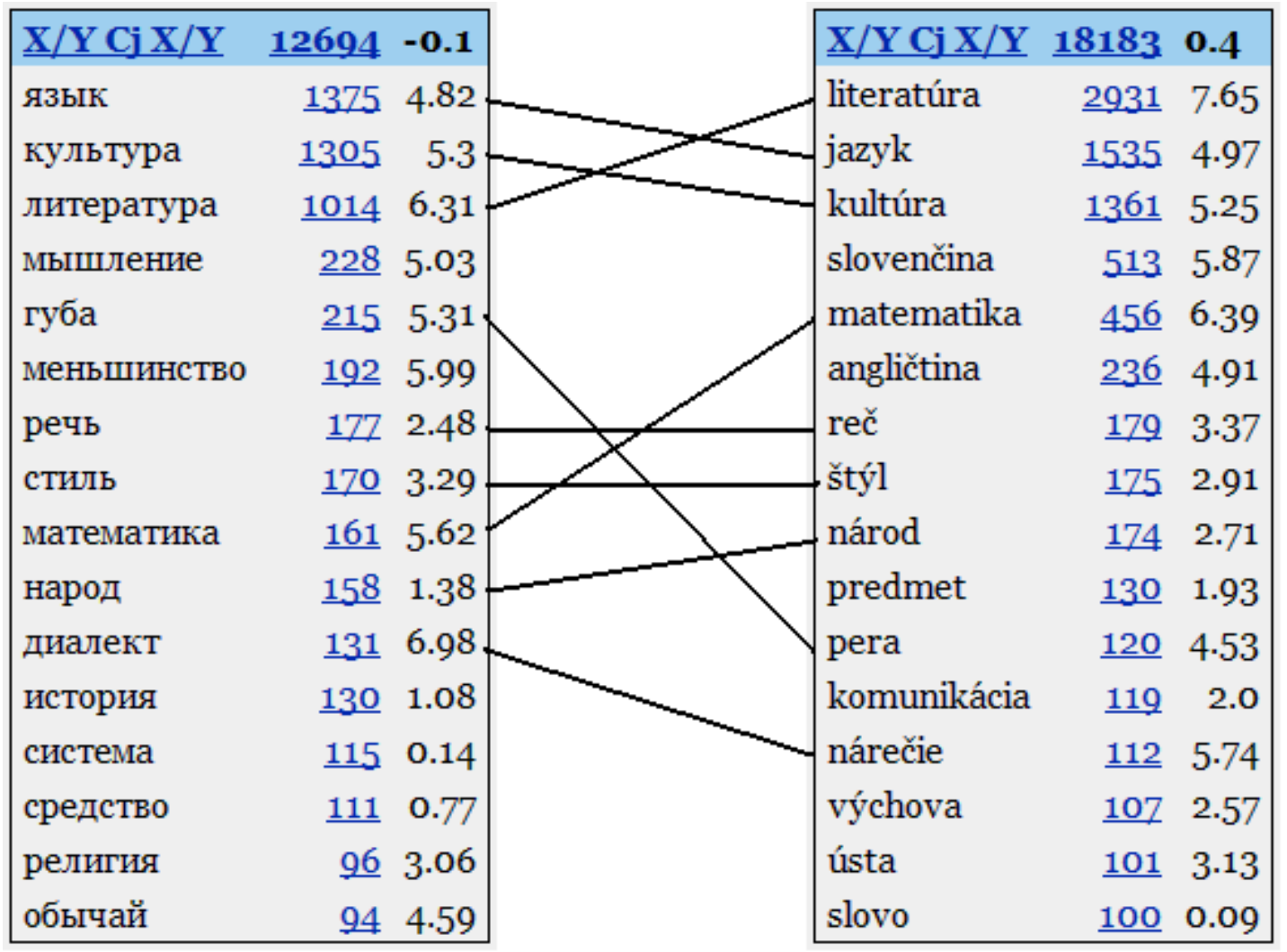


Vb X/X Vb	71035	-0.0
быть	4816	1.81
знать	2806	4.83
говорить	2382	4.22
являться	2025	3.74
владеть	1642	7.64
изучать	1554	7.41
найти	1490	4.82
мочь	1407	1.65
учить	1135	6.86
стать	974	2.42
иметь	870	2.44
выучить	767	7.96
находить	720	5.72
означать	715	5.29
понимать	699	4.18
использовать	640	3.42

Vb X/X Vb	66131	0.1
byť	16560	1.11
ovládať	3010	7.9
mať	2958	1.05
hovoriť	2505	4.21
používať	1734	4.68
učiť	1719	5.99
môcť	1283	0.78
naučiť	1111	5.34
vedieť	861	1.44
musieť	612	0.97
vyučovať	516	6.54
stať	481	1.57
študovať	449	4.95
rozprávať	435	4.27
začať	405	1.27
chcieť	388	0.1

Параллельные скетчи

язык (ru) / jazyk (sk)



X/Y Cj X/Y	12694	-0.1
язык	1375	4.82
культура	1305	5.3
литература	1014	6.31
мышление	228	5.03
губа	215	5.31
меньшинство	192	5.99
речь	177	2.48
стиль	170	3.29
математика	161	5.62
народ	158	1.38
диалект	131	6.98
история	130	1.08
система	115	0.14
средство	111	0.77
религия	96	3.06
обычай	94	4.59

X/Y Cj X/Y	18183	0.4
literatúra	2931	7.65
jazyk	1535	4.97
kultúra	1361	5.25
slovenčina	513	5.87
matematika	456	6.39
angličtina	236	4.91
reč	179	3.37
štýl	175	2.91
národ	174	2.71
predmet	130	1.93
pera	120	4.53
komunikácia	119	2.0
nárečie	112	5.74
výchova	107	2.57
ústa	101	3.13
slovo	100	0.09

Заключение

- **Программа SpiderLing является очень эффективным и удобным средством для создания веб-корпусов**

Заключение

- **Программа SpiderLing является очень эффективным и удобным средством для создания веб-корпусов**
- **Создание совместимых скетч-грамматик осуществимо**

Заключение

- **Программа SpiderLing является очень эффективным и удобным средством для создания веб-корпусов**
- **Создание совместимых скетч-грамматик осуществимо**
- **Параллелизм между скетчами наблюдается также на русском и даже на французском языках**

Заключение

- **Программа SpiderLing является очень эффективным и удобным средством для создания веб-корпусов**
- **Создание совместимых скетч-грамматик осуществимо**
- **Параллелизм между скетчами наблюдается также на русском и даже на французском языках и это любопытно, не правда ли?**

Песочница

http://sketch.juls.savba.sk/sandbox

логин: spb

пароль: spb2013,,

будет действителен по 14 июля 2013

e-mail для справок

vladob@juls.savba.sk