Sketching Words in Comparable Corpora

Vladimír Benko
Slovak Academy of Sciences
Ľ. Štúr Institute of Linguistics
Bratislava

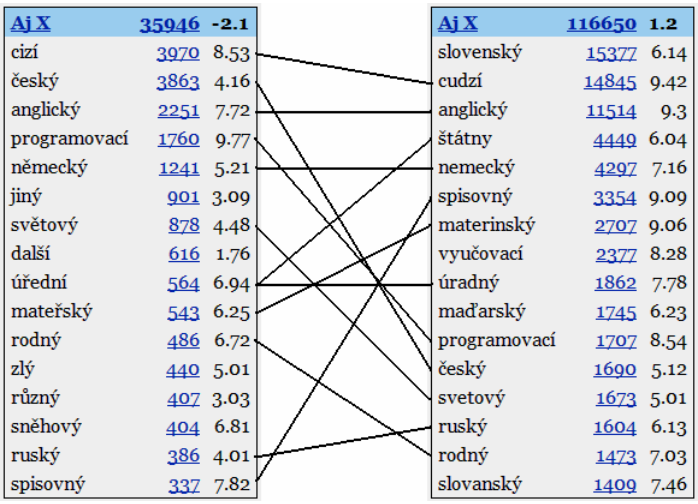
Keywords: word sketches, comparable corpora, bilingual lexicography

Abstract:

Despite the recent advances in building parallel corpora, their current sizes and compositions are still far below the expectations of lexicographers compiling a bilingual dictionary. This is true even for the "big" language pairs, let alone the "smaller" ones, such as Czech and Slovak.

In our Institute, a project of a new two-volume Czech to Slovak dictionary is currently being carried out. As a source of lexical evidence, a 2.34 GW Slovak corpus, a 445 MW Czech corpus, and a 10 MW parallel Czech-Slovak corpus are being used. The monolingual corpora can be accessed via the Word Sketch Engine (installed at our Institute's servers) using compatible word sketch grammars for both Slovak and Czech.

Using dual screen workstations, our lexicographers typically inspect word sketches for a headword in Czech and its translation equivalent(s) in Slovak side by side. They are often amazed how similar the collocational behaviour is at both sides, for great many of the headwords. To illustrate the phenomenon, let us compare the adjectival collocates of the word *jazyk* "language" (it has the same form both in Czech and Slovak) derived from two web corpora:

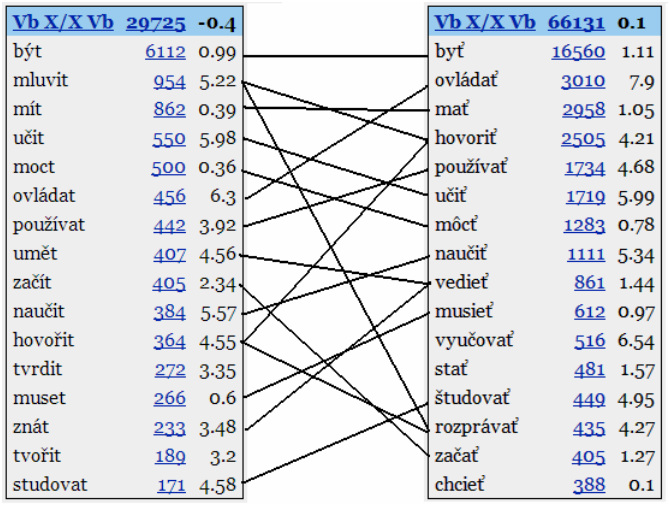


Aj X	35946	-2.1
cizí	3970	8.53
český	3863	4.16
anglický	2251	7.72
programovací	1760	9.77
německý	1241	5.21
jiný	901	3.09
světový	878	4.48
další	616	1.76
úřední	564	6.94
mateřský	543	6.25
rodný	486	6.72
zlý	440	5.01
různý	407	3.03
sněhový	404	6.81
ruský	386	4.01
spisovný	337	7.82

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

It can be seen that out of the 16 most frequent collocates at the Czech side, there are 11 that have their corresponding translation equivalents at the Slovak side (indicated by a line).

Even better situation (14 of 16) can be observed with the verbal collocates:



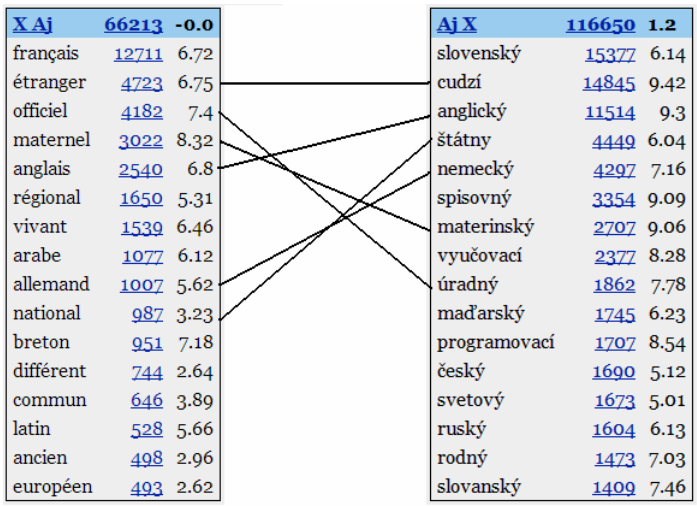
Vb X/X Vb	29725	-0.4
být	6112	0.99
mluvit	954	5.22
mít	862	0.39
učit	550	5.98
moct	500	0.36
ovládat	456	6.3
používat	442	3.92
umět	407	4.56
začít	405	2.34
naučit	384	5.57
hovořit	364	4.55
tvrdit	272	3.35
muset	266	0.6
znát	233	3.48
tvořit	189	3.2
studovat	171	4.58

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

Now, there is an interesting question: is the similarity of word sketches caused simply by the fact that Czech and Slovak are both linguistically and culturally closely related languages? Or, could something like that be observed with other languages as well?

To find an answer, we decided to build two web corpora – a French corpus *Francogallicum* and a Russian corpus *Russicum*. As all the tools necessary were either freely available (SpiderLing, TreeTagger) or could be easily recycled from our own tools, this took only about 3 weeks to complete. After having ported our sketch grammar to French and Russian tagsets, we now have a device to study potential similarities in collocations among all four languages.

Let us have a look at the already mentioned word *language* in French vs. Slovak.

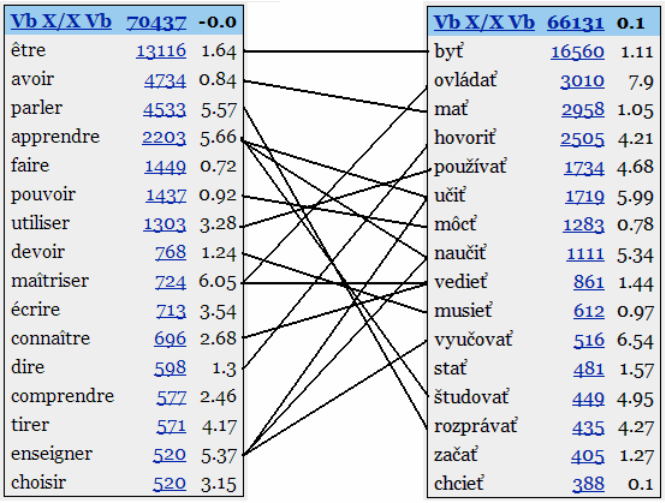


X Aj	66213	-0.0
français	12711	6.72
étranger	4723	6.75
officiel	4182	7.4
maternel	3022	8.32
anglais	2540	6.8
régional	1650	5.31
vivant	1539	6.46
arabe	1077	6.12
allemand	1007	5.62
national	987	3.23
breton	951	7.18
différent	744	2.64
commun	646	3.89
latin	528	5.66
ancien	498	2.96
européen	493	2.62

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

The adjectival collocates have less links than that of Czech vs. Slovak: the cultural difference is apparent here. At the French side we can see adjectives like *French, Arabic, Breton and Latin*, while at the Slovak side there are *Slovak, Hungarian, Czech,* and *Russian* that, in fact, belong to the same lexical group at both sides. Moreover, the leading *French* and *Slovak* are both self-referential. We can safely consider the sketches similar.

And now the verbal collocates.



Vb X/X Vb	70437	-0.0
être	13116	1.64
avoir	4734	0.84
parler	4533	5.57
apprendre	2203	5.66
faire	1449	0.72
pouvoir	1437	0.92
utiliser	1303	3.28
devoir	768	1.24
maîtriser	724	6.05
écrire	713	3.54
connaître	696	2.68
dire	598	1.3
comprendre	577	2.46
tirer	571	4.17
enseigner	520	5.37
choisir	520	3.15

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

The situation is more complex here. This is caused partially by the possible synonyms in translation equivalents at the Slovak side, as well as by the "hidden" difference of *teach* vs. *learn* in Slovak word sketch, as it is expressed analytically (*učiť* and *učiť sa*, respectively). Anyway, the number of links is high.

This small probe demonstrates the overall picture seen among all four languages. The links between sketches typically cover about 50% of keywords at both sides, regardless of the number of lines in the respective tables. Thus, the use of Sketch Engine can unveil a property of languages that could hardly be noticeable without such a tool. This is quite fascinating, isn't it?