Sketching Words in Comparable Corpora

Vladimír Benko

vladob@juls.savba.sk

Slovak Academy of Sciences
Ľ. Štúr Institute of Linguistics

Comenius University in Bratislava
UNESCO Chair for Translation Studies

SKEW 4
Tallinn, 16 October 2013

“Parallel” sketches

jazyk (cs) / jazyk (sk) “language”

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

“Parallel” sketches

jazyk (cs) / jazyk (sk) “language”

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

Aranea

A family of (comparable) web corpora

Motivation: available corpora

- **Do not cover all languages needed**
- **Are of different sizes**
- **Are mostly too big for classroom use**
- **Sketch grammars are too different**

Aranea

Araneum (pl. *aranea*, n.) … **Latin expression for (cob)web**

- **Slovak-Centric (languages spoken and/or taught in Slovakia and the neighbouring countries)**
- **Crawled and pre-processed by *SpiderLing* at (approximately) the same time**
- **Language-independent filtration by the same tools**
- **Language-dependent filtration by the same methodology**

Aranea

- **Compatible tokenization strategy**
- **PoS-tagged by (possibly) free tools (*Tree Tagger*, etc.)**
- **Sentence-segmented**
- **Sentence-level deduplicated, duplicate sentences marked**
- **Word sketches with compatible sketch grammars**
- **“Language-neutral” (Latin) names denoting the language and size**

Aranea

Four sizes for each corpus planned

Maius (greater) … basic version, approx. 1 billion tokens

Minus (smaller) … 10 % sample of Maius (for teaching purposes)

Minumum (minimal) … 1 % sample of Maius (not accessible by general users, used for toolchain and sketch grammar experiments)

Maximum (maximal) … as much as we can get

Aranea

Four Aranea family members available by now

Araneum Russicum Maius & Minus (Russian Web, Version 13.10)

Araneum Francogallicum Maius & Minus (French Web, Version 13.10)

Araneum Germanicum Maius & Minus (German Web, Version 13.10)

Araneum Hispanicum Maius & Minus (Spanish Web, Version 13.11, will be published soon)

Compatible sketch grammars

- **Common set of rules for all languages**
- **Fixed order of tables in word sketches**
- **Rule names represent collocational relationships (i.e. not syntactic)**
- **Syntactic functions of keywords and/or collocates not indicated (e.g., for nouns, we do not speak about "subjects", "objects", or "modifiers", we just indicate the "left-hand" and "right-hand" collocates**

Compatible sketch grammars

- **Word class (PoS) of keyword is not indicated (each rule works for any PoS)**
- **Recall preferred over precision (some tables may not be relevant for all word classes, output may contain lots of noise)**

Compatible sketch grammars

Symbols in rule names

- **X … keyword (of any PoS, except for punctuation)**
- **Y … collocate of any PoS, except for conjunction, preposition and punctuation**
- **Z … collocate of PoS not covered be explicit rules (“catch all” rules)**

Compatible sketch grammars

Symbols in rule names

Nn, Vb, Aj, Av, Pp, Cj … collocate of indicated PoS

***Aa*(X), *aaa*(X) … (in unary rules) PoS categories and subcategories of keyword**

Compatible sketch grammars

Summary of rule names

Binary rules

Nn X; X Nn; Aj X; X Aj; Pp X; X Pp; Z X; X Z … side-sensitive rules

Vb X/X Vb; Av X/X Av … side-insensitive rules

Symmetric rules

X/Y, X/Y; X/Y Cj X/Y … coordination

Compatible sketch grammars

Summary of rule names

Trinary rules

Y Pp X; Y Pp X; Pp Y X; Pp X Y …

prepositions

Unary rules

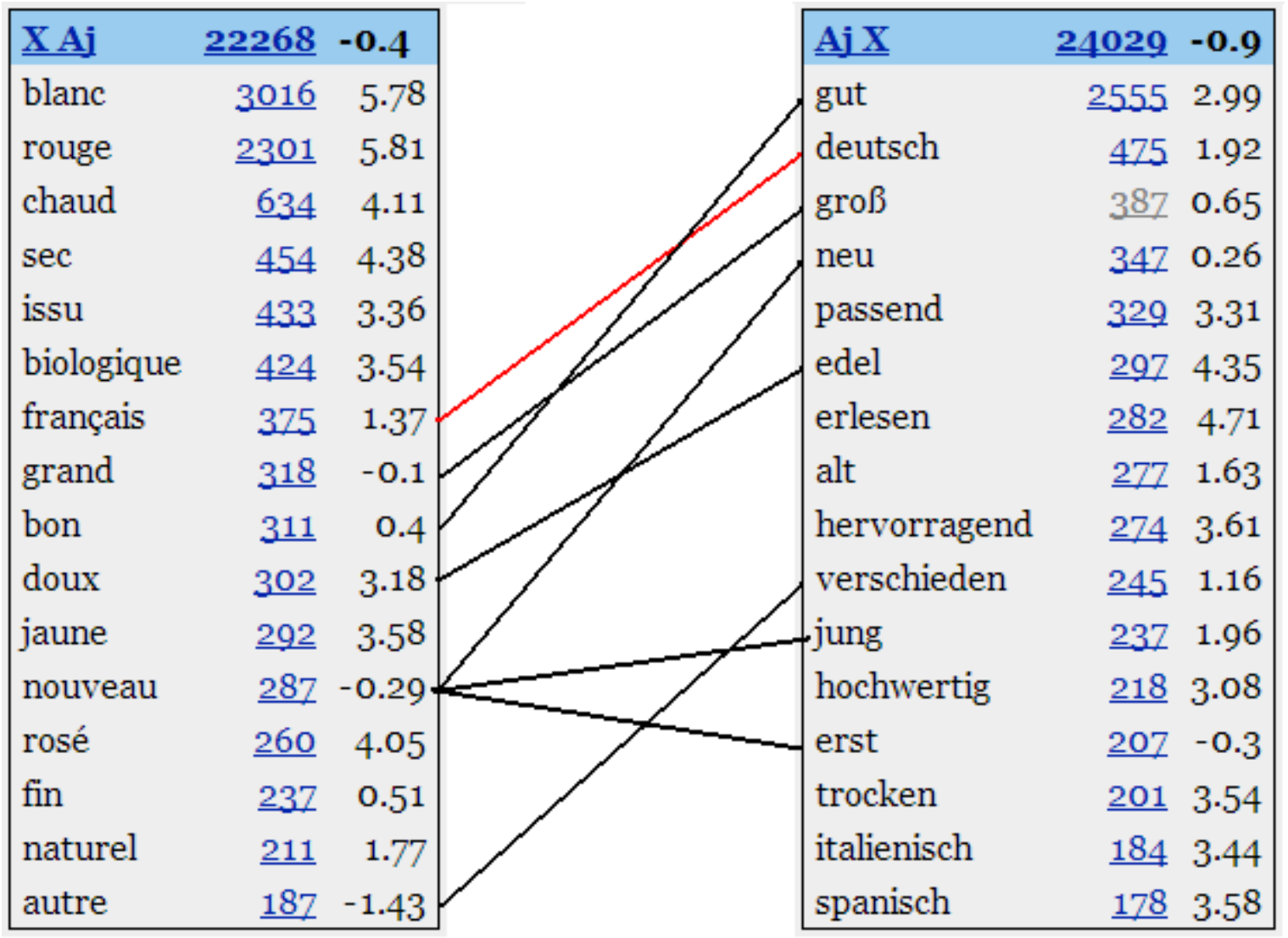
Ar(X); Nn(X); … PoS categories

msc(X); fem(X); … PoS subcategories

No dual rules

“Parallel” sketches

vin (fr) / Wein (de) “wine”

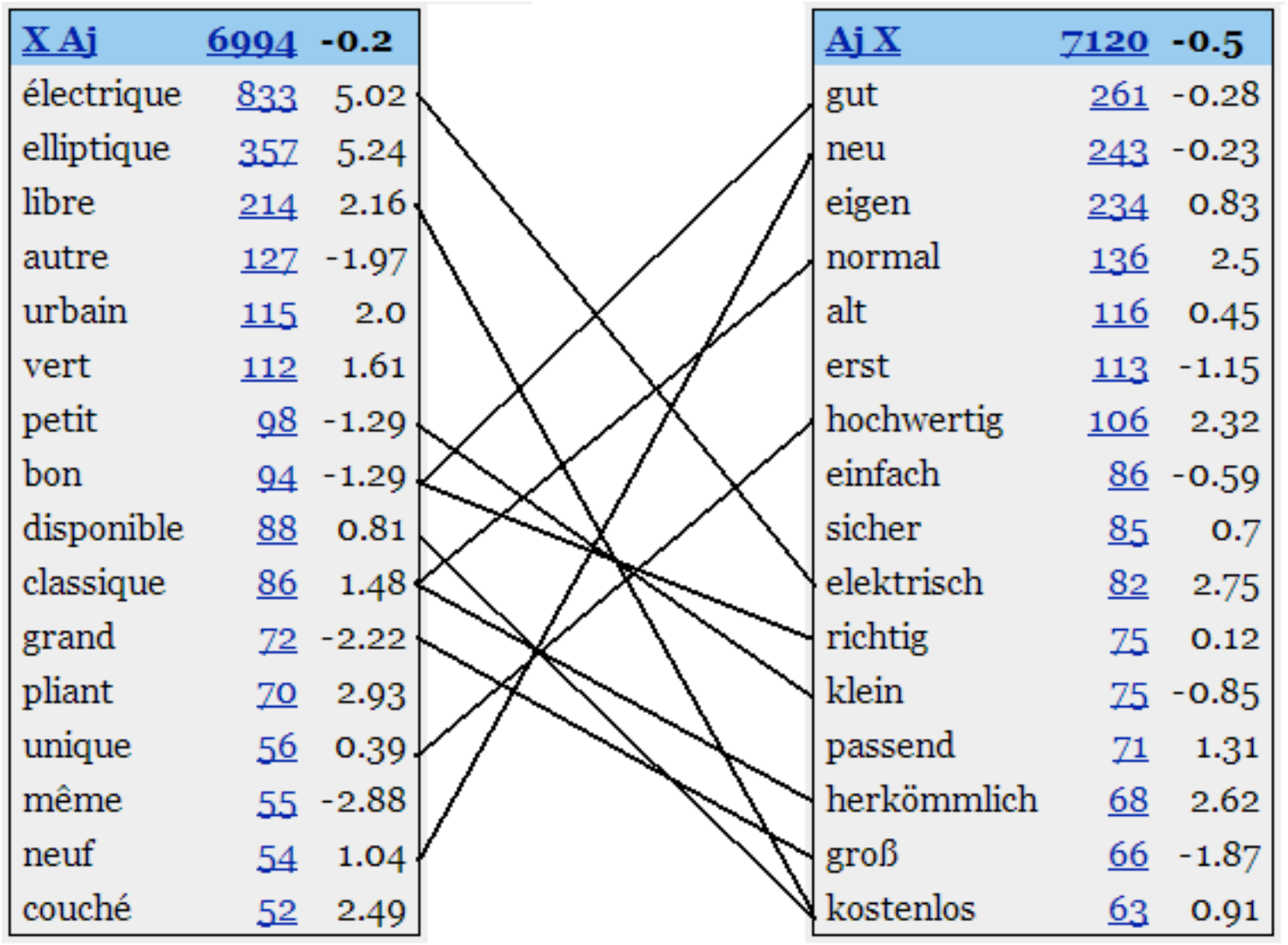


X Aj	22268	-0.4
blanc	3016	5.78
rouge	2301	5.81
chaud	634	4.11
sec	454	4.38
issu	433	3.36
biologique	424	3.54
français	375	1.37
grand	318	-0.1
bon	311	0.4
doux	302	3.18
jaune	292	3.58
nouveau	287	-0.29
rosé	260	4.05
fin	237	0.51
naturel	211	1.77
autre	187	-1.43

Aj X	24029	-0.9
gut	2555	2.99
deutsch	475	1.92
groß	387	0.65
neu	347	0.26
passend	329	3.31
edel	297	4.35
erlesen	282	4.71
alt	277	1.63
hervorragend	274	3.61
verschieden	245	1.16
jung	237	1.96
hochwertig	218	3.08
erst	207	-0.3
trocken	201	3.54
italienisch	184	3.44
spanisch	178	3.58

“Parallel” sketches

vélo (fr) / Rad (de) “bike”

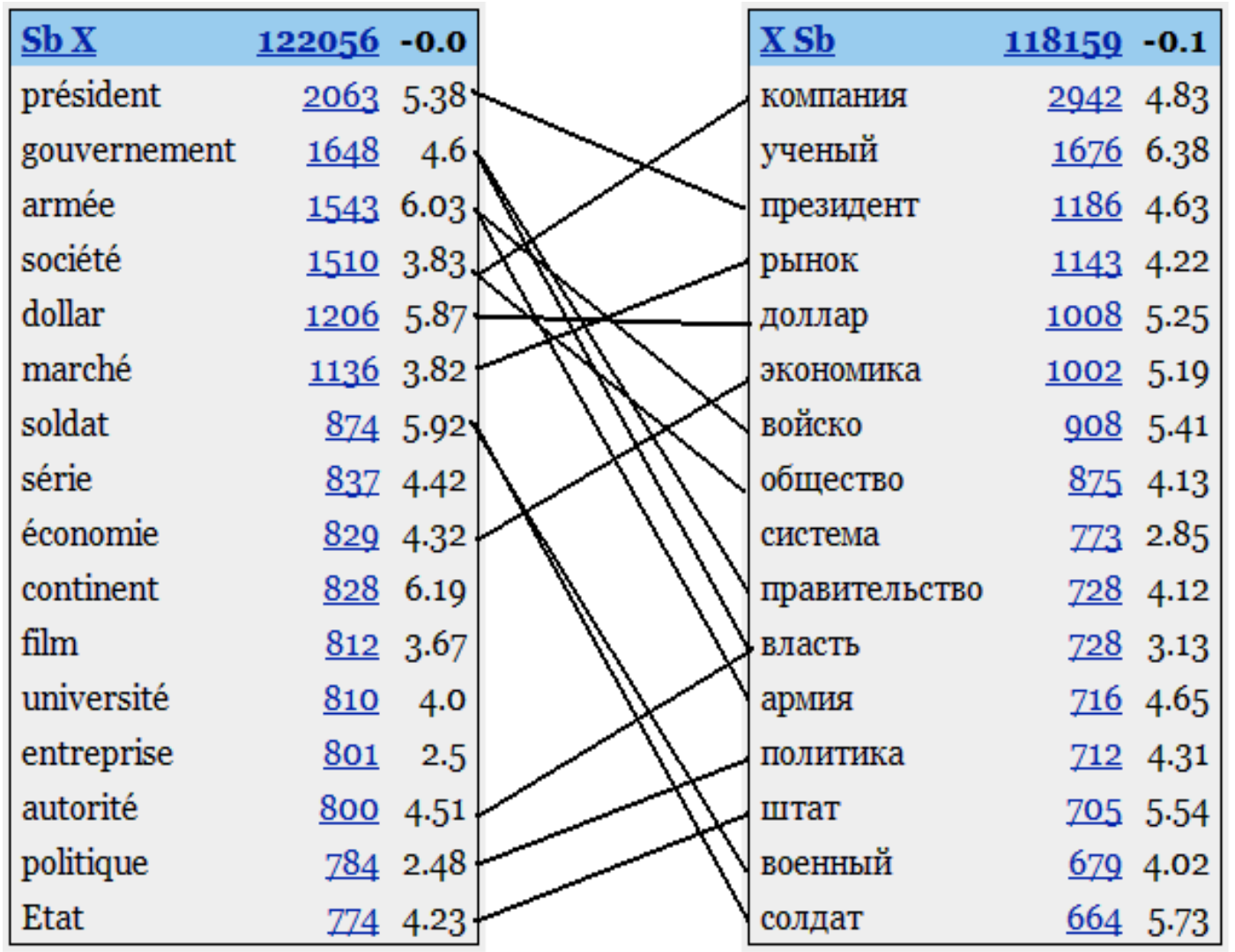


X Aj	6994	-0.2
électrique	833	5.02
elliptique	357	5.24
libre	214	2.16
autre	127	-1.97
urbain	115	2.0
vert	112	1.61
petit	98	-1.29
bon	94	-1.29
disponible	88	0.81
classique	86	1.48
grand	72	-2.22
pliant	70	2.93
unique	56	0.39
même	55	-2.88
neuf	54	1.04
couché	52	2.49

Aj X	7120	-0.5
gut	261	-0.28
neu	243	-0.23
eigen	234	0.83
normal	136	2.5
alt	116	0.45
erst	113	-1.15
hochwertig	106	2.32
einfach	86	-0.59
sicher	85	0.7
elektrisch	82	2.75
richtig	75	0.12
klein	75	-0.85
passend	71	1.31
herkömmlich	68	2.62
groß	66	-1.87
kostenlos	63	0.91

“Parallel” sketches

américain (fr) / американский (ru) “American”

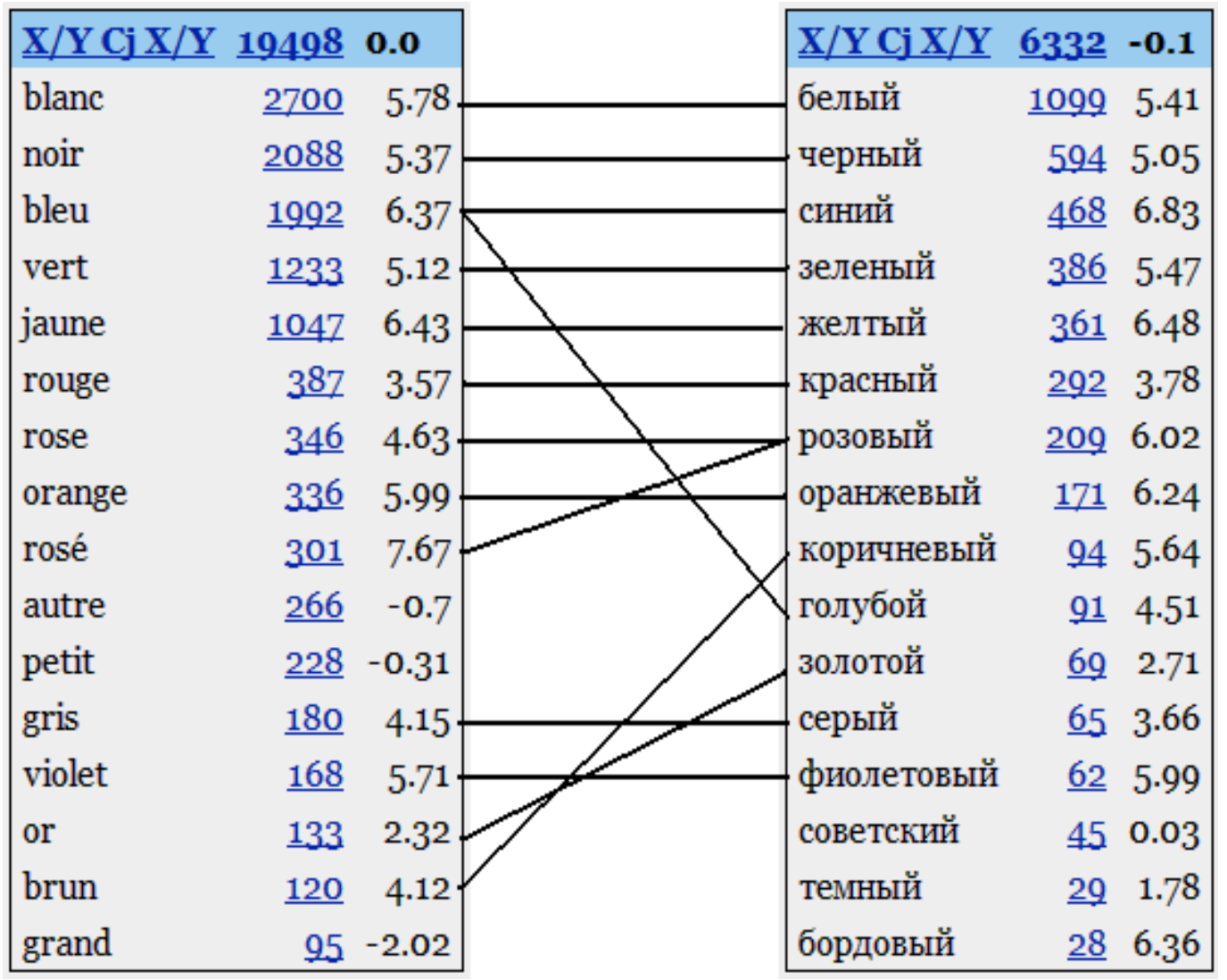


Sb X	122056	-0.0
président	2063	5.38
gouvernement	1648	4.6
armée	1543	6.03
société	1510	3.83
dollar	1206	5.87
marché	1136	3.82
soldat	874	5.92
série	837	4.42
économie	829	4.32
continent	828	6.19
film	812	3.67
université	810	4.0
entreprise	801	2.5
autorité	800	4.51
politique	784	2.48
Etat	774	4.23

X Sb	118159	-0.1
компания	2942	4.83
ученый	1676	6.38
президент	1186	4.63
рынок	1143	4.22
доллар	1008	5.25
экономика	1002	5.19
войско	908	5.41
общество	875	4.13
система	773	2.85
правительство	728	4.12
власть	728	3.13
армия	716	4.65
политика	712	4.31
штат	705	5.54
военный	679	4.02
солдат	664	5.73

“Parallel” sketches

rouge (fr) / красный (ru) “red”



X/Y Cj X/Y	19498	0.0
blanc	2700	5.78
noir	2088	5.37
bleu	1992	6.37
vert	1233	5.12
jaune	1047	6.43
rouge	387	3.57
rose	346	4.63
orange	336	5.99
rosé	301	7.67
autre	266	-0.7
petit	228	-0.31
gris	180	4.15
violet	168	5.71
or	133	2.32
brun	120	4.12
grand	95	-2.02

X/Y Cj X/Y	6332	-0.1
белый	1099	5.41
черный	594	5.05
синий	468	6.83
зеленый	386	5.47
желтый	361	6.48
красный	292	3.78
розовый	209	6.02
оранжевый	171	6.24
коричневый	94	5.64
голубой	91	4.51
золотой	69	2.71
серый	65	3.66
фиолетовый	62	5.99
советский	45	0.03
темный	29	1.78
бордовый	28	6.36

Compatible sketch grammars

Corpora with compatible sketch grammars

- **supplement to (or instead of non-existent) parallel corpora**
- **teaching foreign languages & translation studies**
- **contrastive linguistic studies**
- **collocationally-based sketch grammar as general-purpose grammar and as a convenient starting point for developing language-, tagset- & purpose-specific grammar**