Newly Born Members of the Aranea Family

Vladimír Benko

vladimir.benko@juls.savba.sk

Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
Comenius University, UNESCO Chair in Plurilingual and Multicultural Communication

RASLAN 2018
Karlova Studánka | 8 December 2018

Aranea

- A Project to create a family of web corpora that could be used (primarily) for teaching purposes
- Started in Spring 2013 (after *SpiderLing* has been published)
- Based (mostly) on tools developed at Masaryk University in Brno ("*Brno Pipeline*")
- Complemented by other *FLOSL* tools with minor own contributions
- "Language neutral" (Latin) names:

  araneum, *pl.* **aranea** *n.* > (spider)net
- Similar to ***TenTen*** corpus family, **but…**

Aranea

- **Uniform size**

  Maius (“larger”) … 1 Gword, 1.2 to 1.25 Gtoken

  Minus (“smaller”) … 10% sample of Maius

  Maximum (“largest”) … (for some languages) … as much as can get

- **Uniform PoS representation** (“*atag*”)

  native tagsets mapped to **Araneum Universal Tagset**

- **Explicit information** on success of **morphological lexicon lookup** (“ztag”)

… “Comparable”

Morphosyntactic Annotation

<s>O pár minút lob toho istého hráča znova zastavilo brvno.</s>

Morphosyntactic Annotation

<s>
O
pár
minút
lob
toho
istého
hráča
znova
zastavilo
brvno
<g/>
.
</s>

Morphosyntactic Annotation

<s>		
O	o	Eu4
pár	pár	NUns4
minút	minúta	SSfp2
lob	lob	SSip7
toho	to	PFns2
istého	istý	AAms2x
hráča	hráč	SSms4
znova	znova	Dx
zastavilo	zastaviť	VLdscn+
brvno	brvno	SSns4
<g/>		
.	.	Z.
</s>		

Morphosyntactic Annotation

<s>			
O	o	Eu4	1
pár	pár	NUns4	1
minút	minúta	SSfp2	1
lob	lob	SSip7	0
toho	to	PFns2	1
istého	istý	AAms2x	1
hráča	hráč	SSms4	1
znova	znova	Dx	1
zastavilo	zastaviť	VLdscn+	1
brvno	brvno	SSns4	1
<g/>			
.	.	Z.	1
</s>			

Morphosyntactic Annotation

<s>			
O	o	Eu4	1
pár	pár	NUns4	1
minút	minúta	SSfp2	1
lob	lob	SSip7	0
toho	to	PFns2	1
istého	istý	AAms2x	1
hráča	hráč	SSms4	1
znova	znova	Dx	1
zastavilo	zastaviť	VLdscn+	1
brvno	brvno	SSns4	1
<g/>			
.	.	Z.	1
</s>			

Morphosyntactic Annotation

<s>			
O	o	Eu4	1
pár	pár	NUns4	1
minút	minúta	SSfp2	1
lob	lob	SSip7	0
toho	to	PFns2	1
istého	istý	AAms2x	1
hráča	hráč	SSms4	1
znova	znova	Dx	1
zastavilo	zastaviť	VLdscn+	1
brvno	brvno	SSns4	1
<g/>			
.	.	Z.	1
</s>			

Morphosyntactic Annotation

word	lemma	atag	tag	ztag
<s>				
O	o	Pp	Eu4	1
pár	pár	Nm	NUns4	1
minút	minúta	Nn	SSfp2	1
lob	lob	Nn	SSip7	0
toho	to	Pn	PFns2	1
istého	istý	Aj	AAms2x	1
hráča	hráč	Nn	SSms4	1
znova	znova	Av	Dx	1
zastavilo	zastaviť	Vb	VLdscn+	1
brvno	brvno	Nn	SSns4	1
<g/>				
.	.	Zz	Z.	1
</s>				

Morphosyntactic Annotation

word	lemma	atag	tag	ztag	lemma	tag	prec
<s>							
O	o	Pp	Eu4	1	o	Eu4	19
pár	pár	Nm	NUns4	1	pár	NUns4	35
minút	minúta	Nn	SSfp2	1	minúta	SSfp2	01
lob	lob	Nn	SSip7	0	lobiť	VMdsb+	XX
toho	to	Pn	PFns2	1	ten	PFms4	04
istého	istý	Aj	AAms2x	1	istý	AAms4x	08
hráča	hráč	Nn	SSms4	1	hráč	SSms4	02
znova	znova	Av	Dx	1	znova	Dx	01
zastavilo	zastaviť	Vb	VLdscn+	1	zastaviť	VLdscn+	03
brvno	brvno	Nn	SSns4	1	brvno	SSns1	03
<g/>							
.	.	Zz	Z.	1	Z	Z	01
</s>							

Morphosyntactic Annotation

word	lemma	atag	tag	ztag	lemma	tag	prec
<s>							
O	o	Pp	Eu4	1	o	Eu4	19
pár	pár	Nm	NUns4	1	pár	NUns4	35
minút	minúta	Nn	SSfp2	1	minúta	SSfp2	01
lob	lob	Nn	SSip7	0	lobiť	VMdsb+	XX
toho	to	Pn	PFns2	1	ten	PFms4	04
istého	istý	Aj	AAms2x	1	istý	AAms4x	08
hráča	hráč	Nn	SSms4	1	hráč	SSms4	02
znova	znova	Av	Dx	1	znova	Dx	01
zastavilo	zastaviť	Vb	VLdscn+	1	zastaviť	VLdscn+	03
brvno	brvno	Nn	SSns4	1	brvno	SSns1	03
<g/>							
.	.	Zz	Z.	1	Z	Z	01
</s>							

Morphosyntactic Annotation

word	lemma	atag	tag	ztag	lemma	tag	prec
<s>							
O	o	Pp	Eu4	1	o	Eu4	19
pár	pár	Nm	NUns4	1	pár	NUns4	35
minút	minúta	Nn	SSfp2	1	minúta	SSfp2	01
lob	lob	Nn	SSip7	0	lobiť	VMdsb+	XX
toho	to	Pn	PFns2	1	ten	PFms4	04
istého	istý	Aj	AAms2x	1	istý	AAms4x	08
hráča	hráč	Nn	SSms4	1	hráč	SSms4	02
znova	znova	Av	Dx	1	znova	Dx	01
zastavilo	zastaviť	Vb	VLdscn+	1	zastaviť	VLdscn+	03
brvno	brvno	Nn	SSns4	1	brvno	SSns1	03
<g/>							
.	.	Zz	Z.	1	Z	Z	01
</s>							

Aranea

Accesible via web interface
NoSketch Engine corpus manager

- URL: **http://unesco.uniba.sk** (main site)
  or **http://aranea.juls.savba.sk** (mirror site)
- **Guest access** (100 M corpora, no password)
- **Full Access** (all corpora, free registration required)

Project web page

- **http://unesco.uniba.sk/aranea_about**

Aranea Nova MMXVIII

- **Francogallicum Canadiense** Minus & Medium (Canadian French, 18.03)
- **Georgianum** Minus & Medium (Georgian, 18.08)
- **Dacoromanicum** Minus & Maius (Romanian, 18.09)
- **Estonicum** Minus & Medium (Estonian, 18.09)
- **Lettonicum** Minus & Medium (Latvian, 18.10)
- **Germanicum II** Minus, Maius & Maximum (German, 18.11)
- **Latinum** Parvum (Latin, 18.11)

	Araneum Francogallicum Canadiense Minus (Canadian French, 18.03) 120 M		*Araneum Francogallicum II Minus (French, 17.06) 120 M*		
word	**frequency**	**frequency/mill**	**frequency**	**frequency/mill**	**Score**
UQAM	7,364	61.4	244	2.0	20.6
ASFC	2,787	23.2	28	0.2	19.6
ACIA	2,669	22.2	30	0.2	18.6
Ontario	23,564	196.3	1,217	10.1	17.7
Yukon	5,485	45.7	206	1.7	17.2
canadiennes	14,468	120.6	733	6.1	17.1
Manitoba	6,132	51.1	251	2.1	16.9
Cégep	4,739	39.5	186	1.5	15.9
Canadiens	16,708	139.2	967	8.0	15.5
admissibles	6,544	54.5	314	2.6	15.4
MDN	1,922	16.0	14	0.1	15.2
Canada	162,019	1350.0	10,567	88.0	15.2
assurance-emploi	2,332	19.4	43	0.4	15.0
ritannique	6,363	53.0	316	2.6	14.9
canadienne	27,203	226.7	1,770	14.7	14.5
Saskatchewan	3,675	30.6	155	1.3	13.8
runswick	4,875	40.6	242	2.0	13.8
IRSC	5,441	45.3	288	2.4	13.6
CÉR	1,716	14.3	15	0.1	13.6
Alberta	6,309	52.6	355	3.0	13.5
uébec	8,469	70.6	519	4.3	13.5
canadien	23,409	195.1	1,678	14.0	13.1
Ottawa	15,500	129.2	1,076	9.0	13.1
canadiens	14,222	118.5	994	8.3	12.9
admissibilité	3,585	29.9	172	1.4	12.7
Québec	128,023	1066.7	10,031	83.5	12.6
CRSH	1,709	14.2	29	0.2	12.3
COSEPAC	1,499	12.5	15	0.1	12.0
admissible	4,578	38.1	273	2.3	12.0
REEE	1,871	15.6	47	0.4	11.9

	Araneum Francogallicum Canadiense Minus (Canadian French, 18.03) 120 M		*Araneum Francogallicum II Minus (French, 17.06) 120 M*		
word	**frequency**	**frequency/mill**	**frequency**	**frequency/mill**	**Score**
canadiennes	14,468	120.6	733	6.1	17.1
admissibles	6,544	54.5	314	2.6	15.4
ritannique	6,363	53.0	316	2.6	14.9
canadienne	27,203	226.7	1,770	14.7	14.5
runswick	4,875	40.6	242	2.0	13.8
uébec	8,469	70.6	519	4.3	13.5
canadien	23,409	195.1	1,678	14.0	13.1
canadiens	14,222	118.5	994	8.3	12.9
admissibilité	3,585	29.9	172	1.4	12.7
admissible	4,578	38.1	273	2.3	12.0
québécoise	11,176	93.1	877	7.3	11.3
eſt	1,217	10.1	1	0.0	11.0
québécois	16,242	135.3	1,394	11.6	10.8
habiletés	3,682	30.7	233	1.9	10.8
collégial	2,636	22.0	144	1.2	10.4
québécoises	3,692	30.8	260	2.2	10.0
provinciaux	4,860	40.5	377	3.1	10.0
anada	5,856	48.8	484	4.0	9.9
ivières	3,920	32.7	293	2.4	9.8
ſe	1,060	8.8	1	0.0	9.8
ententes	4,119	34.3	315	2.6	9.8
renseignements	41,026	341.8	4,123	34.3	9.7
aurent	7,050	58.7	638	5.3	9.5
entrevues	3,452	28.8	262	2.2	9.4
directrices	4,959	41.3	430	3.6	9.2
fiducie	1,986	16.5	111	0.9	9.1
autochtones	11,574	96.4	1,181	9.8	9.0
répondants	4,175	34.8	360	3.0	9.0
cosse	2,934	24.4	222	1.8	8.9
provinciale	4,236	35.3	369	3.1	8.9
bienfaisance	3,039	25.3	235	2.0	8.9

	Araneum Francogallicum Canadiense Minus (Canadian French, 18.03) 120 M		*Araneum Francogallicum II Minus (French, 17.06) 120 M*		
word	**frequency**	**frequency/mill**	**frequency**	**frequency/mill**	**Score**
canadiennes	14,468	120.6	733	6.1	17.1
admissibles	6,544	54.5	314	2.6	15.4
ritannique	6,363	53.0	316	2.6	14.9
canadienne	27,203	226.7	1,770	14.7	14.5
runswick	4,875	40.6	242	2.0	13.8
uébec	8,469	70.6	519	4.3	13.5
canadien	23,409	195.1	1,678	14.0	13.1
canadiens	14,222	118.5	994	8.3	12.9
admissibilité	3,585	29.9	172	1.4	12.7
admissible	4,578	38.1	273	2.3	12.0
québécoise	11,176	93.1	877	7.3	11.3
eſt	1,217	10.1	1	0.0	11.0
québécois	16,242	135.3	1,394	11.6	10.8
habiletés	3,682	30.7	233	1.9	10.8
collégial	2,636	22.0	144	1.2	10.4
québécoises	3,692	30.8	260	2.2	10.0
provinciaux	4,860	40.5	377	3.1	10.0
anada	5,856	48.8	484	4.0	9.9
ivières	3,920	32.7	293	2.4	9.8
ſe	1,060	8.8	1	0.0	9.8
ententes	4,119	34.3	315	2.6	9.8
renseignements	41,026	341.8	4,123	34.3	9.7
aurent	7,050	58.7	638	5.3	9.5
entrevues	3,452	28.8	262	2.2	9.4
directrices	4,959	41.3	430	3.6	9.2
fiducie	1,986	16.5	111	0.9	9.1
autochtones	11,574	96.4	1,181	9.8	9.0
répondants	4,175	34.8	360	3.0	9.0
cosse	2,934	24.4	222	1.8	8.9
provinciale	4,236	35.3	369	3.1	8.9
bienfaisance	3,039	25.3	235	2.0	8.9

Araneum Georgianum

Trdelnik
SOUVENIRS
CHIMNEY CAKE
CHIMNEY CAKE
ICE CREAM
FRESH JUICE
SANGRIA
COFFEE

Araneum Georgianum

Araneum Georgianum

New challenges

- The amount of Georgian web data is rather small
- Georgian script does not distinguish letter case, i.e., sentence splitter procedure had to be modified
- The only available PoS Tagger is based on a rather small (10 K) morphological lexicon covering just some 75% of corpus tokens
- To be able to query (alternatively) in Latin script as well, a transliteration was needed

Araneum Georgianum

Solutions

- Multiple independent crawls with more seed URLs yielded to some 250 M tokens
- Good luck here: Georgian language does not use full-stop final abbreviations extensively
- TnT with the available language model far from ideal, yet acceptable
- Both word form and lemma transliterated (as *rword* and *rlemma*, respectively), *Simple query* CQL expression modified to include them both

Araneum Georgianum

ghvino ▾ Araneum Georgianum Medium (Georgian, 18.08) 254 M

Query **ghvino** **50,973** (200.92 per million)

Page 1 of 2,549 Go Next | Last

24saati.ge	; საქართველოს საელჩო გამოფენას ქართული	ღვინით	მოამარაგებს და კოლექციის გადაზიდვაში
jesus.ge	დასახლდებიან; ჩაყრიან ვენახებს და დალევენ მათ	ღვინოს	, გააშენებენ ბაღებს და შეჭამენ მათ ხილს.
article.wn.com	მომსახურებით: კულტურული, ისტორიული და	ღვინის	ტურებით, ასევე ექსტრემალური და
ycpdgeo.wordpress.com	. საქართველოდან რა უნდა წაგველო თუ არა	ღვინო	(ჩურჩხელებიც გვქონდა) □. ¶ ≡ მართლაც რომ ენით
saqartvelo.biz	ონლაინში ქართულად და სხვა მრავალი. ¶ ≡ სიახლეები ¶	ღვინის	ბიზნესი მსოფლიო ბაზარზე ¶ 2014-01-18 ¶ ქართული
saqartvelo.biz	ბიზნესი მსოფლიო ბაზარზე ¶ 2014-01-18 ¶ ქართული	ღვინო	მსოფლიო ბაზრებზე რომ ნაკლებადაა წარმოდგენილი
investinsamegrelo.ge	წარმოება, მარტვილში - ჩაის გადამუშავება და	ღვინის	წარმოება, სენაკში - ღვინის წარმოება, ქ. ფოთში -
investinsamegrelo.ge	გადამუშავება და ღვინის წარმოება, სენაკში -	ღვინის	წარმოება, ქ. ფოთში - თევზის ზეთის, თევზის
e2e4paiki.wordpress.com	ღვინოშია ჭეშმარიტებაო – ამბობენ. რატომაც არა,	ღვინო	მაგარი რამ არის! კარგად იცით, საახალწლოდ
natakhtari.wordpress.com	დამახასიათებელია ხილისა და პარფიუმერიის,	ღვინის	ტიპის არომატები. აქედან გამომდინარე, მიიღება
gfsis.org	დიდი აქცენტის გაკეთება არ შეიძლება. მხოლოდ	ღვინოსა	და ჩაიზე არ უნდა ვფიქრობდეთ.
club-monadire.ge	, ბელელ -, თიბ-, თეს-, სკა, ყანა, ცელ-, ვენაჯ,	ღვინო	, სელ-, საწნეხელ-, ფოცხ-, ღერღ- (ღერღილი),
gmmstudio.com.ge	ყოველწლიურად ტარდება ქალაქ თბილისში. ¶ ≡	ღვინის	ტური ¶ რელიგიური ტური ¶ აქტიური ტურები ¶
bpi.ge	(IFC) გამოყო, 23 მლნ... ¶ თბილისი (BPI) -	ღვინის	ეროვნული სააგენტო უძველეს მარანში
bpi.ge	. მარანი 1962 წელს, საქართველოში გამართული	ღვინის	მსოფლიო კონგრესის ფარგლებში აშენდა. მარანში
bpi.ge	, სადაც სხვადასხვა სახეობის 25 000 ბოთლი	ღვინო	ინახება, მათ შორის ყველაზე ძველი ღვინოც,
evpatori.ge	მარადიულია. ¶ ავტორი: admin ¶ ქართული	ღვინო	ისტორიით ერთ-ერთი უძველესია ევროპაში.
evpatori.ge	ყველაზე ადრეული ნიშნები მევენახეობისა და	ღვინის	წარმოებაზე მიუთითებს საქართველოში ღვინის
evpatori.ge	და ღვინის წარმოებაზე მიუთითებს საქართველოში	ღვინის	კულტურის არსებობაზე დაახ. 7000 წლის წინ, რაც
evpatori.ge	დაახ. 7000 წლის წინ, რაც საქართველოს	ღვინის	სამშობლოს სტატუსს ანიჭებს. ¶ ავტორი: admin ¶

Page 1 of 2,549 Go Next | Last

	Cooccurrence count	Candidate count	T-score	MI	logDice
წითელი	1,397	30,667	37.211	7.824	9.131
ქართული	4,475	278,324	66.059	6.322	8.798
ქარხანა	721	12,001	26.761	8.224	8.551
ქვევრი	525	2,584	22.890	9.981	8.327
ზოთლი	476	5,529	21.766	8.743	8.108
დაყენების	437	2,063	20.884	10.042	8.076
კლუბი	671	43,991	25.562	6.246	7.855
თეთრი	602	35,100	24.248	6.415	7.840
წარმოება	546	33,763	23.076	6.330	7.722
მარანი	318	2,818	17.800	9.133	7.597
კახური	298	2,043	17.238	9.503	7.525
ჭიქა	305	7,600	17.376	7.641	7.414
მწარმოებელი	284	5,363	16.788	8.042	7.367
ექსპორტი	273	3,752	16.477	8.500	7.352
ტურიზმი	307	14,187	17.358	6.750	7.270
ხარისხი	457	49,815	20.909	5.512	7.215
ფესტივალი	304	18,503	17.222	6.353	7.163
ეროვნული	832	140,929	27.862	4.876	7.150
დაყენებული	226	2,856	14.995	8.621	7.104
ცქრიალა	204	528	14.275	10.909	7.020
ექსპორტის	208	4,729	14.356	7.774	6.935
მშრალი	211	5,748	14.446	7.513	6.929
სუფრა	220	11,141	14.681	6.618	6.858
ნატურალური	182	3,789	13.434	7.901	6.766
დეგუსტაცია	166	499	12.876	10.693	6.723
ქვევრში	167	847	12.909	9.938	6.722
დასახელების	175	3,374	13.177	8.012	6.721
კომპანია	649	151,781	24.278	4.411	6.712
დაყენება	172	3,537	13.060	7.919	6.692
ტკბილი	163	6,838	12.659	6.890	6.529

	Cooccurrence count	Candidate count	T-score	MY	logDice
Red	1,397	30,667	37.211	7.824	9.131
Georgian	4,475	278,324	66.059	6.322	8.798
Factory	721	12,001	26.761	8.224	8.551
pitcher	525	2,584	22.890	9.981	8.327
The bottle	476	5,529	21.766	8.743	8.108
Installing	437	2,063	20.884	10.042	8.076
Club	671	43,991	25.562	6.246	7.855
White	602	35,100	24.248	6.415	7.840
Production	546	33,763	23.076	6.330	7.722
Cellar	318	2,818	17.800	9.133	7.597
Kakhetian	298	2,043	17.238	9.503	7.525
A glass	305	7,600	17.376	7.641	7.414
Manufacturer	284	5,363	16.788	8.042	7.367
Export	273	3,752	16.477	8.500	7.352
Tourism	307	14,187	17.358	6.750	7.270
quality	457	49,815	20.909	5.512	7.215
Festival	304	18,503	17.222	6.353	7.163
National	832	140,929	27.862	4.876	7.150
Installed	226	2,856	14.995	8.621	7.104
Sparkling	204	528	14.275	10.909	7.020
Export	208	4,729	14.356	7.774	6.935
Dry	211	5,748	14.446	7.513	6.929
Supra	220	11,141	14.681	6.618	6.858
Natural	182	3,789	13.434	7.901	6.766
Tasting	166	499	12.876	10.693	6.723
In qvevri	167	847	12.909	9.938	6.722
Denomination	175	3,374	13.177	8.012	6.721
Company	649	151,781	24.278	4.411	6.712
Set up	172	3,537	13.060	7.919	6.692
Sweet	163	6,838	12.659	6.890	6.529