Language Code Switching
In Web Corpora

Vladimír Benko

vladimir.benko@juls.savba.sk

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

RASLAN 2017
Karlova Studánka, 1–3 December 2017

The Lexicographers' Perspective

Sources of lexical evidence

(a) "Traditional" Corpora

Texts **covered by copyright** and received from the respective copyright owners

Text types

- Fiction (novels, short stories, poetry, …)
- Non-fiction (academic writings, memoires, travelogues, …)
- Media (newspapers, magazines, …)

Pre-defined **sampling strategy**, only "quality" texts accepted

Foreign-language texts / text samples rarely encountered

Usually small or **medium-sized**

The Lexicographers' Perspective

Sources of lexical evidence

(b) Web Corpora

Texts downloaded from the Internet by means of automated procedures, **copyright status often not clear**

Text types

- Presentations of institutions (companies, schools, …)
- Electronic media
- E-commerce
- Blogs, social media, discussion forums
- Unclassified texts stored in data clouds / archives
- (Almost) no fiction

Sampling strategy **difficult to apply**

The Lexicographers' Perspective

Sources of lexical evidence

(b) Web Corpora

Texts (**mostly**) **not proofread even by the author**,
i.e., quality rather low, lots of "**noise**"

- Mistakes and errors, non-standard orthography
- Texts without diacritics (or partial use of diacritics)
- Emotions expressed by typography (character repetition, all-caps, emoticons/emoji)
- **Language code switching** (e.g., Czech/Slovak & Ukrainian/Russian in discussions)
- Usually **very large** (billions of tokens)

Web Corpora

How much is 1 billion words?

The Bible's about 611,000 words long in its original languages.

That's about 3 times the length of *Moby Dick*, 35% longer than *The Lord of the Rings*, and about 24,000 more words than *War and Peace*.

How many words are there in the Bible?

The King James Authorized Bible has 783,137 words. How many words is that? If you can type at 60 words a minute, it would take you just over 217 and a half hours to retype the entire Bible. Can you imagine how long type-setting must have taken in the early days of printing?

Web Corpora

How much is 1 billion words?

200 words per minute

12,000 words per hour

96,000 words per (working) day

Web Corpora

How much is 1 billion words?

200 words per minute

12,000 words per hour

96,000 words per (working) day

Working days during a year (approximately)**:**

365 – 105 – 10 – 30 = 365 – 145 **= 220**

21,220,00 words per year

Web Corpora

How much is 1 billion words?

200 words per minute

12,000 words per hour

96,000 words per (working) day

Working days during a year (approximately)**:**

365 – 105 – 10 – 30 = 365 – 145 **= 220**

21,220,00 words per year

Years in working life

62 – 16 = 46

971,520,000 words during working life

Lexicographers' Perspective

Why to use web corpora in lexicography

- Can be (usually much) **larger** than traditional corpora, i.e., *more suitable for analysis of infrequent phenomena, such as phraseology*
- New text types, genres, domains & registers, *larger proportion of more informal language*
- Shorter development/update cycle, i.e, new language phenomena and tendencies can be identified earlier: *lexical neologisms, adaptation of loanwords*

Ideal solution (if applicable) – merging (largest) traditional with the web corpus for the respective language

prim-6.1-all & *Araneum Slovacum Maximum* … **Omnia Slovaca** (4.49 G)

syn v4 & *Araneum Bohemicum Maximum* … **Omnia Bohemica** (9.53 G)

Lexicographers' Perspective

Problem in using web and "merged" corpora

- **Language code switching**, i.e., foreign language text fragments in otherwise monolingual text
- In all Aranea corpora large proportion of **English text fragments** (including in Chinese & Arabic corpora)

In Slovak corpora

- **Czech** text fragments (usually in discussions, e-shops, etc.)
- Text **fragments without diacritics** (both Slovak and Czech)
- **English** text fragments

Language Code Switching

Language identification within the "Brno Pipeline" (integrated in *SpiderLing*)

Language components (with separate language models)

- ***chared*** (web page encoding detection) … *based on characters/trigrams?…* language model supplied
- ***jusText*** (boilerplate removal) … *based on list of stop words* … supplied
- ***trigrams*** (webpage languge identification) … *based on character trigrams* (*unigrams for CJK and some other languages*) … language model created based on a text sample supplied by the user

Output of *SpiderLing* is fairly good but language identification fails in

(1) distinguishing very similar languages, (2) identifying very short texts, (3) texts with language code switching

Language Code Switching

Why lexicographers do / have to bother:

- Inter-lingual (“false”) homographs in concordances and word sketches:

pr-web.sk	píská a šumí, mozek se neprokrvuje... Z toho **plyne** špatná paměť, zapomínání, přidá se vysoký	☑
beo.sk	ti amici jazdit ked nie na ruskej rope a **plyne** ...na kravske prdy? ¶ _____	☐
pluska.sk	peniaze. Ady má totiž domček v Lozorne, kde na **plyne** varí, ohrieva vodu i kúri. A každé usporené	☐
pluska.sk	je golfové ihrisko. Takže čo usporím na **plyne** , vrazím do golfu,“ vysvetľuje. Pritom ale	☐
euroekonom...	od roku 2001. Kupříkladu, z dat UAH MSU **plyne** ochlazovací trend mezi lednem 2001 a květnem	☑
euroekonom...	sopkami a ENSO-m) ¶ Co z vývoje Ap indexu **plyne** ? Nejspíš to, že nás čeká další rok extrémně	☑
zahori.est...	Francúzsko. K odstráneniu závislosti na ruskom **plyne** môžu prispieť aj náleziská plynu na južnom	☑
zahori.est...	troch až piatich rokov bude Slovensko na **plyne** nezávislé“. ¶ Expremiér Mikuláš Dzurinda	☐
despitebor...	nariadenie nasledovala smernica o zemnom **plyne** v roku 1998 (98/30/EC). V oboch sa požadovalo	☐
despitebor...	2003/55/EC (5) o elektrickej energii a o **plyne** predstavujú najväčšie zmeny doterajšieho	☐
despitebor...	znamenalo vyvlastnenie, čo by hlavne pri zemnom **plyne** viedlo k zvýšeniu cien pre koncových zákazníkov	☐
burjanosko...	střední školy a jeho tři zástupci. ¶ - Co z toho **plyne** : měli bychom si, jako daňoví poplatníci	☑
energia.sk	zaostávame. V Čechách mení v elektrine a zemnom **plyne** dodávateľa 15 až 20 percent domácností.	☐
energia.sk	plyn ho dokáže nahradiť, dopyt po zemnom **plyne** bude rásť a cena pôjde tiež hore. ¶ Ak sa	☐
diskusie.s...	pripravili na vojnu, a pred začiatkom klamstva o **plyne** boli v Izraeli hlavní americkí generáli	☐
sixpack.cz	telekomu atd.. Prakticky aj pri elektrike, **plyne** a vode. Tam neexistuje alternatíva, ale	☐

jen Araneum Slovacum Quadruplex II ter (Slovak, 17.11) 194 M freq = 2,141 (11.03 per million)

usage patterns		
Nn(X)	360	16.81
Xy(X)	1,781	83.19

X/Y , X/Y		
	212	9.90
protože	5	8.32
jak	7	7.97
co	14	7.81
se	9	7.72

X/Y Cj X/Y		
	84	3.92
frank	6	10.33
dolár	7	8.43
euro	6	5.73

Y X		
	1,044	48.76
přece	23	9.34
přeci	16	8.92
japonský	53	8.23
nejsou	9	7.89
jsem	10	5.77
nebo	13	5.59
dolár	7	5.26
posilniť	7	5.16
miliarda	6	4.55
spať	6	4.51
uz	7	4.00
stačiť	17	3.87
zase	5	3.00
milión	5	2.60
asi	8	1.72
už	32	0.99

X Y		
	1,132	52.87
kvůli	9	7.92
Selter	8	7.83
těžko	6	7.38
jsem	23	7.06
několik	5	6.94
stúpnuť	7	6.15
posilniť	5	4.76
pár	20	3.35
trochu	9	3.11
škoda	5	2.54
málo	9	2.35
jeden	47	2.16
malý	12	1.27
asi	5	1.07
otázka	5	1.04
5	5	0.94

Nn X		
	894	41.76
přece	23	9.34
přeci	17	9.14
nejsou	9	7.84
můžu	5	7.42
jsem	42	6.77
kterou	5	6.62
nebo	32	6.11
dolár	15	5.81
euro	6	1.76
kurz	5	1.17
rok	6	-2.06

X Nn		
	1,142	53.34
kvůli	10	7.92
Selter	8	7.81
několik	6	6.95
část	5	6.89
jsem	27	6.73
nebo	14	3.95
škoda	11	2.94
tím	6	0.84
pomoc	7	0.83
otázka	9	0.76
pocit	5	0.58
čas	7	-1.00
rok	7	-2.51

Aj X		
	149	6.96
japonský	54	8.27
možný	7	0.03

X Aj		
	328	15.32
samotný	5	1.09
malý	15	0.88
posledný	5	0.36
dobrý	5	-1.04

Vb X/X Vb		
	1,467	68.52
teď	13	6.78
zatím	7	6.68
Teď	5	6.46
chybieť	5	6.28
jsme	14	5.88
oslabiť	11	5.62
stúpnuť	9	4.14
posilniť	16	3.91
klesnúť	8	2.85
spať	6	2.10
stačiť	23	1.90
držať	5	0.29

Av X/X Av		
	120	5.60
trochu	9	1.78
stále	5	-0.59

Z X		
	1,763	82.34
ne	15	7.38
pak	20	7.32
vlastně	10	7.29
není	19	7.13
prostě	10	7.07
já	10	7.04
se	89	7.01
jsou	23	6.98
může	10	6.85
jde	7	6.81
třeba	9	6.68
je	30	6.63

X Z		
	1,787	83.47
proto	23	8.46
dál	21	8.35
jednou	10	7.18
něco	11	7.04
doporučit	7	6.96
na	34	6.85
nevím	6	6.70
pro	36	6.53
když	8	6.38
jejich	7	6.28
takové	5	6.23
před	5	6.22

Language Code Switching

Why lexicographers do / have to bother:

- False items in lists of words not recognized by morphological annotation, i.e., potential neologisms

4064	18	zameriavaní	zameriavanie	zameriavanie		Nn
4065	6	zameriavaný	zameriavaný	zameriavaný		Xx
4066	3	zamerikanizovaný	zamerikanizovaný	zamerikanizovaný		Xx
4067	3	zámernejšie	zámernejšie	zámerný	aj	Nn
4068	3	záměrně	záměrně	záměrně	yy	Xy
4069	4	zamerom	zameor	zámer		Nn
4070	3	zamerov	zameor	zámer		Nn
4071	3	záměrům	záměrům	záměrům	yy	Xy
4072	3	zameru	zameru	zámer		Nn
4073	12	záměru	záměru	záměru	yy	Xy
4074	5	zamery	zamery	zámer		Nn
4075	4	zamery	zamery	zámer	nn	Xy
4076	7	záměry	záměry	záměry	yy	Xy
4077	6	záměr	záměr	záměr	yy	Xy
4078	3	zameškaného	zameškaný	zameškaný		Nn
4079	7	zameškaného	zameškaný	zameškaný		Xx
4080	201	zameškané	zameškaný	zameškaný		Xx
4081	15	zameškanú	zameškaný	zameškaný		Xx

Language Code Switching

The task

- **Identify** "wrong" text fragments in the (possibly) lowest reasonable level: our decision … sentence level
- **Delete** or **mark** them so that they not interfere with "good" text in concordances, word sketches and word lists
- Possibly **enable** some **evaluation** of the whole procedure

Language Code Switching

The task

- Identify “wrong” text fragments in the (possibly) lost reasonable level: our decision … sentence level
- Delete of mark them so that they not interfere with “good” text in concordances, word sketches and word lists
- Possibly enable some evaluation of the whole procedure

Language identification is considered “solved”:

- Character/n-gram-based approach
- Dictionary-based approach

But: both usually fails for (a) short texts, (2) similar languages, (3) mixed language content

Language Code Switching

Proposed solution

- Dictionary approach with very large (“exhaustive”) dictionaries
- Standard tools used for PoS tagging:

  Slovak and *Slovak without diacritics* … *TreeTagger with own language model* (SNK tagset)

  Czech … *MorphoDiTa with the newest Czech language model* (Hajič tagset)

  English … *TreeTagger with standard language model* (Penn tagset)
- Merge annotations
- Decide on every wordform
- Use summary information to decide on sentence

Language Code Switching

```
<s>
Skladba             skladba             Nn SSfs1    1
vyšla               vyjsť               Vb VLdscf+  1
16                  @card@              Nb 0        1
.                   Zz                  Z. Z.       1
septembra           september           Nn SSis2    1
prostredníctvom     prostredníctvom     Pp Eu2      1
Motown              Motown              Xy %        0
Records             records             Xy %        1
a                   a                   Cj O        1
podiela             podiela             Vb VLdsaf+  0
na                  na                  Pp Eu6      1
nej                 ona                 Pn PFfs6    1
aj                  aj                  Pt T        1
s                   s                   Pp Eu7      1
americkým           americký            Aj AAms7x   1
rapperom            rapperom            Nn SSms7    0
Juicy               Juicy               Nn SSip4    0
J-om                J-om                Yy Q        0
.                   .                   Zz Z.       1
<s>
```

Language Code Switching

```
<s>
Skladba          skladba          Nn SSfs1   1  Nn 1
vyšla            vyjsť            Vb VLdscf+ 1  Vb 1
16               @card@           Nb 0       1  Nb 1
.                Zz               Z. Z.      1  Z. 1
septembra        september        Nn SSis2   1  Nn 1
prostredníctvom  prostredníctvom  Pp Eu2     1  Pp 1
Motown           Motown           Xy %       0  Ab 0
Records          records          Xy %       1  Xy 1
a                a                Cj O       1  Cj 1
podiela          podiela          Vb VLdsaf+ 0  Vb 1
na               na               Pp Eu6     1  Pp 1
nej              ona              Pn PFfs6   1  Pn 1
aj               aj               Pt T       1  Pt 1
s                s                Pp Eu7     1  Pp 1
americkým        americký         Aj AAms7x  1  Aj 1
rapperom         rapperom         Nn SSms7   0  Nn 0
Juicy            Juicy            Nn SSip4   0  Nn 0
J-om             J-om             Yy Q       0  Yy 0
.                .                Zz Z.      1  Zz 1
<s>
```

Language Code Switching

```
<s>
Skladba          skladba          Nn SSfs1   1  Nn 1  Nn 1
vyšla            vyjsť            Vb VLdscf+ 1  Vb 1  Vb 1
16               @card@           Nb 0       1  Nb 1  Nm 1
.                Zz               Z. Z.      1  Zz 1  Zz 1
septembra        september        Nn SSis2   1  Nn 1  Yy 0
prostredníctvom  prostredníctvom  Pp Eu2     1  Pp 1  Yy 0
Motown           Motown           Xy %       0  Ab 0  Nn 1
Records          records          Xy %       1  Xy 1  Yy 0
a                a                Cj O       1  Cj 1  Aj 1
podiela          podiela          Vb VLdsaf+ 0  Vb 1  Yy 0
na               na               Pp Eu6     1  Pp 1  Ij 1
nej              ona              Pn PFfs6   1  Pn 1  Aj 1
aj               aj               Pt T       1  Pt 1  Aj 1
s                s                Pp Eu7     1  Pp 1  Nn 1
americkým        americký         Aj AAms7x  1  Aj 1  Aj 1
rapperom         rapperom         Nn SSms7   0  Nn 0  Yy 0
Juicy            Juicy            Nn SSip4   0  Nn 0  Yy 0
J-om             J-om             Yy Q       0  Yy 0  Yy 0
.                .                Zz Z.      1  Zz 1  Zz 1
<s>
```

Language Code Switching

```
<s lang="Sk.1|sk8:4|9:5|cs6:3|en3:0" dtokens="13" stokens="19">
Skladba          skladba          Nn SSfs1   1  Nn 1  Nn 1  Nn NP   0
vyšla            vyjsť            Vb VLdscf+ 1  Vb 1  Vb 1  Nn NP   0
16               @card@           Nb 0       1  Nb 1  Nm 1  Nm CD   1
.                Zz               Z. Z.      1  Zz 1  Zz 1  Zz SENT 1
septembra        september        Nn SSis2   1  Nn 1  Yy 0  Nn NN   0
prostredníctvom  prostredníctvom  Pp Eu2     1  Pp 1  Yy 0  Nn NN   0
Motown           Motown           Xy %       0  Ab 0  Nn 1  Nn NP   1
Records          records          Xy %       1  Xy 1  Yy 0  Nn NPS  1
a                a                Cj O       1  Cj 1  Aj 1  Dt DT   1
podiela          podiela          Vb VLdsaf+ 0  Vb 1  Yy 0  Nn NN   0
na               na               Pp Eu6     1  Pp 1  Ij 1  Xy TO   1
nej              ona              Pn PFfs6   1  Pn 1  Aj 1  Nn NP   1
aj               aj               Pt T       1  Pt 1  Aj 1  Nn NP   0
s                s                Pp Eu7     1  Pp 1  Nn 1  Vb VVZ  0
americkým        americký         Aj AAms7x  1  Aj 1  Aj 1  Nn NN   0
rapperom         rapperom         Nn SSms7   0  Nn 0  Yy 0  Nn NN   0
Juicy            Juicy            Nn SSip4   0  Nn 0  Yy 0  Aj JJ   1
J-om             J-om             Yy Q       0  Yy 0  Yy 0  Nn NN   0
.                .                Zz Z.      1  Zz 1  Zz 1  Zz SENT 1
<s>
```

jen Araneum Slovacum Quadruplex II ter (Slovak, 17.12) 194 M freq = 523 (2.69 per million)

jen Araneum Sl

Click on collocates to access reciprocal bilingual search

usage patterns			usage patterns		
Nn(X)	214	40.92	Nn(X)	360	16.81
Xy(X)	309	59.08	Xy(X)	1,781	83.19

X/Y Cj X/Y			X/Y Cj X/Y		
	34	6.50		84	3.92
frank	6	10.91	frank	6	10.33
dolár	7	8.54	dolár	7	8.43
euro	6	5.76	euro	6	5.73

X/Y , X/Y		
	212	9.90
protože	5	8.32
jak	7	7.97
co	14	7.81
se	9	7.72

Y X			Y X		
	339	64.82		1,044	48.76
japonský	52	8.40	přece	23	9.34
bilión	4	7.35	přeci	16	8.92
dolár	7	5.44	japonský	53	8.23
posilniť	7	5.33	nejsou	9	7.89
miliarda	6	4.68	jsem	10	5.77
uz	6	3.86	nebo	13	5.59
milión	5	2.65	dolár	7	5.26
stačiť	4	1.81	posilniť	7	5.16
ďalej	4	1.15	miliarda	6	4.55
nie	9	-0.24	spať	6	4.51
@card@	10	-2.15	uz	7	4.00
mať	5	-2.81	stačiť	17	3.87
byť	21	-3.18	zase	5	3.00
sa	4	-4.79	milión	5	2.60
			asi	8	1.72
			už	32	0.99

X Y			X Y		
	297	56.79		1,132	52.87
Selter	8	9.72	kvůli	9	7.92
stúpnuť	7	6.58	Selter	8	7.83
posilniť	5	4.97	těžko	6	7.38
dolár	4	4.34	jsem	23	7.06
jeden	10	-0.06	několik	5	6.94
len	4	-1.29	stúpnuť	7	6.15
@card@	13	-2.32	posilniť	5	4.76
sa	9	-3.39	pár	20	3.35
byť	7	-4.50	trochu	9	3.11
			škoda	5	2.54
			málo	9	2.35
			jeden	47	2.16
			malý	12	1.27
			asi	5	1.07
			otázka	5	1.04
			5	5	0.94

X Aj		
	328	15.32
samotný	5	1.09
malý	15	0.88
posledný	5	0.36
dobrý	5	-1.04

Nn X			Nn X		
	207	39.58		894	41.76
dolár	15	5.93	přece	23	9.34
mena	4	4.23	přeci	17	9.14
euro	6	1.78	nejsou	9	7.84
kurz	5	1.19	můžu	5	7.42
			jsem	42	6.77

X Nn			X Nn		
	247	47.23		1,142	53.34
Selter	8	9.89	kvůli	10	7.92
dolár	4	2.94	Selter	8	7.81
pondelok	4	2.50	několik	6	6.95
			část	5	6.89
			jsem	27	6.73

Language Code Switching

Problems

- The Czech tagger is “too good”
- The English tagger tends to tag short strings as NP (proper noun), regardless of capitalization
- Very short sentences are difficult to decide

Ideas for improvement

- Improve the Slovak tagger (enlarge the lexicon)
- Use “worse” Czech tagger, i.e., one with smaller lexicon
- Use PoS information (such as number of finite verbs) as a parameter in decision