university of groningen

Predicting code-switching in Udmurt/Russian blogs

Masha Medvedeva
s3026817

Methodology & Statistics

ReMa Language and Communication Technologies

2015/2016

Code-mixing

Zentella: I'll tell you exactly when I have to leave, at ten o'clock. Y son las nueve y cuarto. ("And it's nine fifteen.")

Marta: Lolita, te voy a dejar con Ana. ("I'm going to leave you with Ana.") Thank you, Ana.

Code-mixing

> 4 000 000 000 bilinguals, trilinguals, etc

'Linguistic rubbish' or an area of linguistic study?

Types of code-mixing

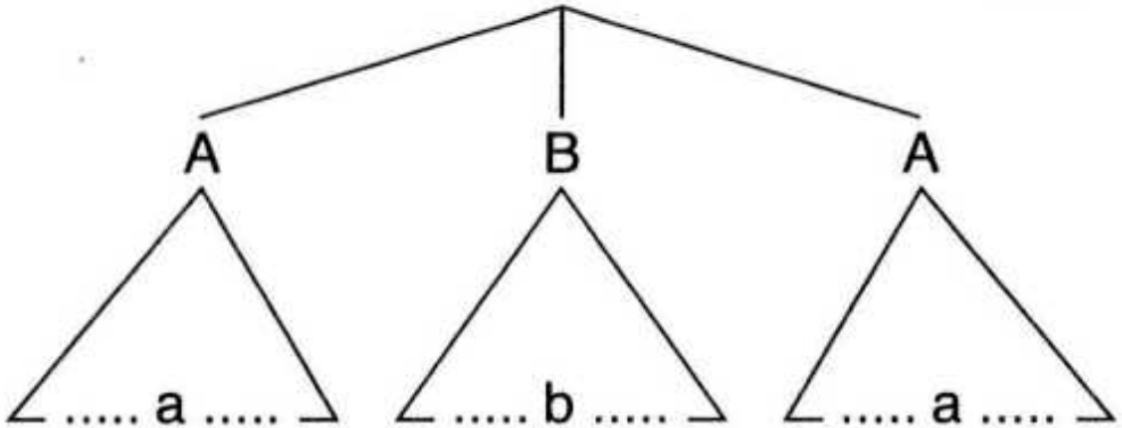
- insertion

- alternation

- congruent lexicalization

Insertion

(11) **insertion**



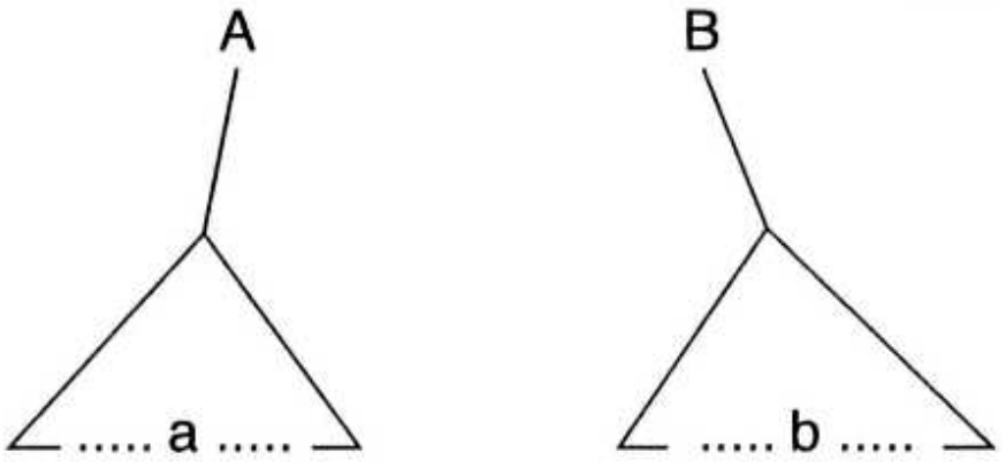
Chay-ta *las dos de la noche*-ta chaya-mu-yku.

that-AC the Iwo of the night-AC arrive-CIS-I pl

'There at two in the morning we arrive.

Alternation

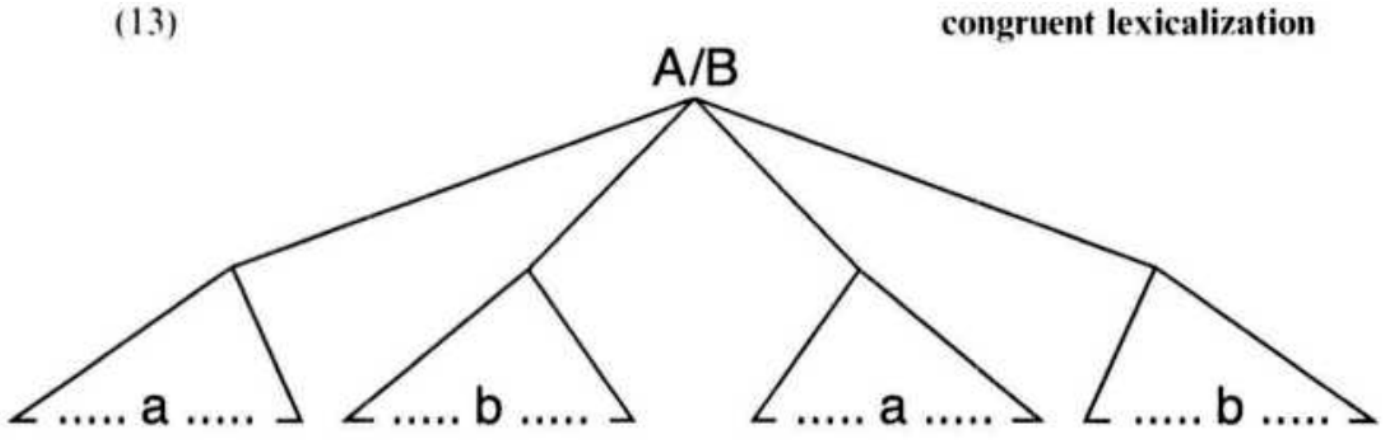
(12) **alternation**



Sometimes I'll start a sentence in Spanish *y termino en español.*

'Sometimes I'll start a sentence in Spanish *and finish in Spanish.'*

Congruent lexicaliztation



wan heri **gedeelte** de ondro **beheer** fu **gewapende machten** one wholepart cop
under control of armed force
“One whole part is under control of the armed forces.”

Dutch–Sranan (Bolle 1994:75, cited in Muysken 2000:139)

Free-Morpheme Constraint

*EAT - iendo

That's what *Papschi mein* -s to say.

'That's what Papschi means to say.'

…in meine Mutter *-s car*.

'In my mother's car.'

Equivalence Constraint

El **MAN** que **CAME** ayer **WANTS JOHN** comprar **A CAR** nuevo.

The	man	who	came	yesterday	wants …	John	to	buy	a	new car.
|	|	|	|	|	|	**X**		|	|	**X**
El	hombre	que	vino	ayer	quire…	que John		compre	un	coche nuevo

Equivalence constraint has been verified as a tendency in many language pairs:

Spanish/English (Poplack 1980), *Finnish/English* (Poplack et al. 1987), *French/Arabic* (Naït M'Barek & Sankoff 1988), *English/Tamil* (Sankoff et al. 1990), *Wolof/French* and *Fongbe/French* (Poplack & Meechen 1995)*, Ukranian/English* (Budzhak-Jones 1995), *French/English* (Turpin 1998) and possibly more.

MAIN

Udmurt/Russian Code Switching corpus

The page you are currently viewing is a web interface for the pilot version of the Udmurt/Russian Code Switching corpus. The corpus size so far is about 146,000 tokens. The texts are taken from blogs written in Udmurt with occasional switching to Russian. The corpus has been automatically annotated with Udmurt and Russian morphological analyzers and contain additional annotation related to code-switching. The texts are also available in the Udmurt corpus.

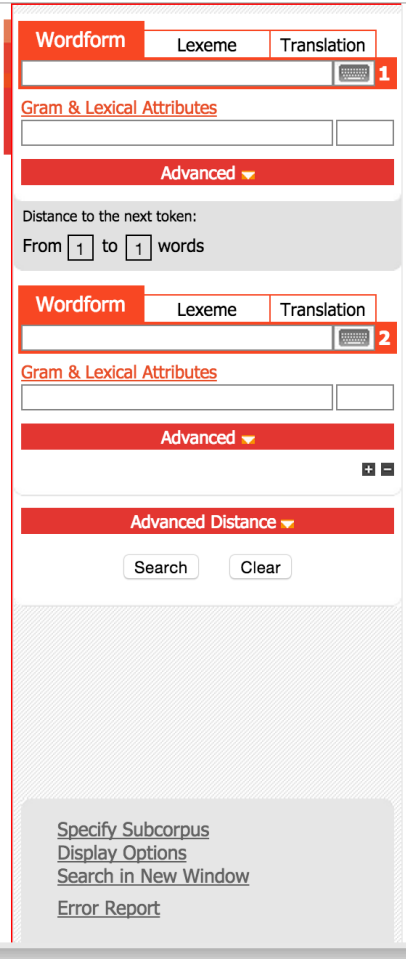
Participants

Maria Medvedeva
Timofey Arkhangelskiy
(HSE School of Linguistics)

Web interface

The search platform of the Eastern Armenian National Corpus (EANC) was used for this corpus. You can read about making search queries at EANC help page.

http://web-corpora.net/UdmurtRussianCorpus/search/



powered by Corpus Technologies

Exact Loose QuickSearch

Data

- Udmurt/Russian

- 7 blogs
- 146,000 tokens
- 18967 sentences
- 5615 sentences contain code-mixing (30%)

- EANC platform
- mystem (Russian), UniParser (Udmurt)

Corpus Examples: Insertion

Пуко пиосмуртъёс юыса ужзы бере ***и*** ми отын ӟабыльтйськом удмурт сяин.

‘Men are sitting, drinking before work ***and*** meanwhile we are there talking in Udmurt.

Атае ***третий десяток пошёл*** шуыса шоккетйз.

‘~Father **is in his thirties now**, that’s what they say.’

Corpus Examples: Congruent Lexicalization

Окно - **со *стекло*** прозрачное***, адзиськод, мар луэ со сьӧрын,*** а чтобы **лэсьтыны сое** *зеркало и чтобы* **адзыны астэ гинэ** *и не замечать*, **мар луэ** *вокруг* **стеклоез** *покрыть***тоно сереброен.**

A window **has a** transparent **glass**, **you can see through it what's going on,** and if you want to **make a** mirror **out of it**, to **see just yourself**, and not notice **what's** around, **the glass has to be** covered **with silver**.

Annotation

8	15	rus		вал	3		2	вал		N	inan,m,nom,norm,sg	.	eos
8	15	rus		вал	3		3	вал		N	acc,inan,m,norm,sg	.	eos
9	1	udm	cap	Но	5	1	1	но	но	CNJ		bos	
9	1	udm	cap	Но	5	1	2	но	и	PART		bos	
9	1	rus	cap	Но	5		3	Но		N	anim,m,nom,norm,persn,sg		bos
9	1	rus	cap	Но	5		4	но		INTJ	norm	bos	
9	1	rus	cap	Но	5		5	но		CONJ	norm	bos	
9	2	udm		малпаськыса		2	2	1	малпаны	1. мыслить, подумать 2. мечтать, помечтать 3. сообразить	V	II pass cnv preced	
9	2	udm		малпаськыса		2	2	2	малпаськыны	1. подумать, думать 2. мечтать, помечтать 3. размышлять	V	I cnv preced	
9	3	udm		мон	1	1	1	мон	я	PRO	nom sg	,	
9	4	udm		карыны	1	1	1	карыны	1.сделать, делать	V	I inf		
9	5	udm		вуи	1	1	1	вуыны	прийти, созреть, попасть, успеть	V	I 1 sg past		
9	6	udm		со	3	1	1	со	он/она/оно	PRO	nom sg		
9	6	rus		со	3		2	сей		APRO	acc,anom,f,sg		
9	6	rus		со	3		3	со		PR	norm		
9	7	udm		малпанэз	2	1	1	малпан	мысль, идея, сознание	N	sg acc	.	eos
9	7	udm		малпанэз	2	1	2	малпан	мысль, идея, сознание	N	sg nom 3sg	.	eos
10	1	udm	cap	Озьысен	1	1	1	озьысен	итак, значит, стало быть	N	sg nom	bos	
10	2	udm		выль	1	1	1	выль	новость	N	sg nom	ins	1
10	3	rus		сайт	2		1	сайт		N	inan,m,nom,norm,sg		
10	3	rus		сайт	2		2	сайт		N	acc,inan,m,norm,sg		
10	4	udm		кылдӥз	1	1	1	кылдыны	1. возникнуть, возникать 2. довестись, случиться	V	I 3 sg pres		
10	5	udm		удмурт	2	1	1	удмурт	удмурт	N	sg nom		
10	5	rus		удмурт	2		2	удмурт		N	anim,m,nom,norm,sg		
10	6	udm		вотэсын	2	1	1	вотэс	паутина, тенёта обл. 3. сеть, сетка	N	sg ine	.	eos
10	6	udm		вотэсын	2	1	2	вотэс	паутина, тенёта обл. 3. сеть, сетка	N	sg ins	.	eos
11	1	udm	cap	Кыкетӥ	1	1	1	кык	два	NUM	sg prol	bos	congr
11	2	udm		доменэн	1	1	1						
11	3	udm		та	2	1	1	та	тот	PRO	sg nom		
11	3	rus		та	2		2	тот		APRO	f,nom,norm,sg		
11	4	rus		сайт	2		1	сайт		N	inan,m,nom,norm,sg		
11	4	rus		сайт	2		2	сайт		N	acc,inan,m,norm,sg		
11	5	udm		ужа	6	1	1	ужаны	1. поработать, работать 2. возделать, возделывать	V	II 3 sg pres	,	
11	5	udm		ужа	6	1	2	ужаны	1. поработать, работать 2. возделать, возделывать	V	II 3 sg pres neg	,	
11	5	udm		ужа	6	1	3	ужаны	1. поработать, работать 2. возделать, возделывать	V	II imper sg subj	,	
11	5	udm		ужа	6	1	4	ужаны	1. поработать, работать 2. возделать, возделывать	V	II sg fut neg	,	
11	5	rus		ужа	6		5	уж		N	anim,gen,m,norm,sg	,	
11	5	rus		ужа	6		6	уж		N	acc,anim,m,norm,sg	,	
11	6	udm		кудӥз	1	1	1					,	
11	7	udm		нимаське	8	2	1	ниманы	назвать, называть, наименовать, именовать	V	II 1 pl pres neg	«	
11	7	udm		нимаське	8	2	2	ниманы	назвать, называть, наименовать, именовать	V	II 2 pl pres neg	«	
11	7	udm		нимаське	8	2	3	ниманы	назвать, называть, наименовать, именовать	V	II pass 3 sg pres	«	
11	7	udm		нимаське	8	2	4	ниманы	назвать, называть, наименовать, именовать	V	II pass imper pl subj	«	
11	7	udm		нимаське	8	2	5	ниманы	назвать, называть, наименовать, именовать	V	II pass pl fut neg	«	
11	7	udm		нимаське	8	2	6	нимаськыны	назваться, называться, наименоваться, именоваться	V	I 3 sg pres	«	

Information Retrieval

sentN	wordN	word	lang	trigger	switchTyp	switch0	POS
2	5	удмурт	trigger	1	ins	1	N
2	6	дискотек	rus	0	ins	1	N
3	7	дуре	trigger	1	ins	1	N
6	13	то	trigger	1	ins	1	CONJ
13	2	вылын	trigger	1	ins	1	ADV
13	3	футболк	rus	0	ins	1	N
15	11	бусы	trigger	1	ins	1	N
16	2	умме	rus	0	ins	1	UNK
17	5	пати	rus	0	ins	1	N
19	2	оля	rus	0	ins	1	N
34	3	пус	rus	0	ins	1	UNK
34	4	кулэ	trigger	1	ins	1	V
34	4	валатэк	trigger	1	ins	1	ADV
36	2	вообще	rus	0	ins	1	PARENTH

Additional annotation: congruent lexicalization

86	10	историк	udm		0	UNK
86	11	малпанзы	udm		0	N
87	1	малпанъ	trigger	congr	0	N
87	2	бугарски	udm		0	UNK
87	3	ява	rus		1	N
87	4	куинь	udm		1	NUM
87	5	дано	trigger		0	V
87	6	адямиос	udm		0	N
87	7	музеязы	trigger		0	N
87	8	вуылйм	trigger		0	V
87	9	на	trigger		0	PR
87	10	цветаева	trigger		0	N
87	11	шишкин	rus		1	N
87	12	дурова	trigger		0	N

Triggering: Chi-square

```
> table(trigger, switch0)
       switch0
trigger     0     1
      0 34612  1505
      1 28752   630
> chisq.test(table(trigger, switch0))

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(trigger, switch0)
X-squared = 209.6, df = 1, p-value < 2.2e-16
```

Logistic Regression

- binary dependent variable
- categorical independent variables
- not normally distributed

Logistic Regression

```
Call:
glm(formula = switch0 ~ trigger, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2918  -0.2918  -0.2918  -0.2082   2.7721

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.13541    0.02633 -119.08   <2e-16 ***
trigger     -0.68534    0.04812  -14.24   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 18818  on 65498  degrees of freedom
Residual deviance: 18600  on 65497  degrees of freedom
AIC: 18604

Number of Fisher Scoring iterations: 6
```

Logistic Regression

```
Call:
glm(formula = switch0 ~ POS, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6945  -0.2700  -0.2700  -0.2166   3.1970

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.001811   0.160007 -18.760  < 2e-16 ***
POSADJ       -1.687496   0.371127  -4.547 5.44e-06 ***
POSADV       -0.995192   0.218497  -4.555 5.25e-06 ***
POSADVPRO     0.699226   0.758685   0.922 0.356722
POSANUM     -10.564255 535.411192  -0.020 0.984258
POSAPRO      -0.555440   0.306872  -1.810 0.070295 .
POSCNJ       -1.930262   0.304549  -6.338 2.33e-10 ***
POSCONJ      -0.738530   0.192627  -3.834 0.000126 ***
POSIMIT     -10.564255 161.432621  -0.065 0.947823
POSINTJ      -1.288648   0.602911  -2.137 0.032567 *
POSINTRJ    -10.564255 138.242662  -0.076 0.939086
POSN         -0.291560   0.164528  -1.772 0.076377 .
POSNPRO      -0.735858   0.530617  -1.387 0.165504
POSNUM        0.761290   0.188132   4.047 5.20e-05 ***
POSPARENTH    1.392373   0.283104   4.918 8.73e-07 ***
POSPART      -0.624067   0.212944  -2.931 0.003382 **
POSPOST      -1.058632   0.305644  -3.464 0.000533 ***
POSPR         0.008785   0.199483   0.044 0.964872
POSPRAEDIC    1.702528   0.670705   2.538 0.011136 *
POSPRO       -2.102562   0.250727  -8.386  < 2e-16 ***
POSUNK        0.135984   0.165585   0.821 0.411512
POSV         -0.723997   0.169863  -4.262 2.02e-05 ***
POSnorm      16.567878 535.411192   0.031 0.975314

    Null deviance: 18818  on 65498  degrees of freedom
Residual deviance: 18188  on 65476  degrees of freedom
AIC: 18234
```

Logistic Regression

```
Call:
glm(formula = switch0 ~ trigger + POS, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8380  -0.3328  -0.2159  -0.1679   3.4879

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.097770   0.168327 -12.462  < 2e-16 ***
trigger      -1.013414   0.059801 -16.946  < 2e-16 ***
POSADJ       -2.522421   0.374524  -6.735 1.64e-11 ***
POSADV       -1.520623   0.221814  -6.855 7.11e-12 ***
POSADVPRO     0.290763   0.767028   0.379 0.704630
POSANUM     -11.468297 535.411195  -0.021 0.982911
POSAPRO      -0.476598   0.307355  -1.551 0.120987
POSCNJ       -2.677258   0.308176  -8.687  < 2e-16 ***
POSCONJ      -0.636167   0.193254  -3.292 0.000995 ***
POSIMIT     -11.256841 159.777042  -0.070 0.943833
POSINTJ      -1.209730   0.603266  -2.005 0.044931 *
POSINTRJ    -11.419283 137.854950  -0.083 0.933982
POSN         -0.761782   0.168083  -4.532 5.84e-06 ***

    Null deviance: 18818  on 65498  degrees of freedom
Residual deviance: 17884  on 65475  degrees of freedom
AIC: 17932

POSNPRO      -0.738395   0.531599  -1.389 0.164830
POSNUM        0.562971   0.189674   2.968 0.002996 **
POSPARENTH    0.979778   0.288075   3.401 0.000671 ***
POSPART      -1.366386   0.218055  -6.266 3.70e-10 ***
POSPOST      -1.862565   0.309625  -6.016 1.79e-09 ***
POSPR        -0.008138   0.200163  -0.041 0.967571
POSPRAEDIC    1.231856   0.684428   1.800 0.071887 .
POSPRO       -2.969338   0.255873 -11.605  < 2e-16 ***
POSUNK       -0.768057   0.173638  -4.423 9.72e-06 ***
POSV         -1.143372   0.172869  -6.614 3.74e-11 ***
POSnorm      15.663836 535.411195   0.029 0.976661
```

Logistic Regression: Triggering vs POS

```
> anova(data.glm0, data.glm1, test="Chisq")
Analysis of Deviance Table

Model 1: switch0 ~ trigger
Model 2: switch0 ~ POS
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     65497      18600
2     65476      18188 21   412.13 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Logistic Regression: POS vs Triggering + POS

```
> anova(data.glm1, data.glm2, test="Chisq")
Analysis of Deviance Table

Model 1: switch0 ~ POS
Model 2: switch0 ~ trigger + POS
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     65476      18188
2     65475      17884  1    304.1 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Logistic Regression: Triggering vs Triggering + POS

```
> anova(data.glm0, data.glm2, test="Chisq")
Analysis of Deviance Table

Model 1: switch0 ~ trigger
Model 2: switch0 ~ trigger + POS
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     65497      18600
2     65475      17884 22   716.23 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Best Model

```
Call:
glm(formula = switch0 ~ trigger + POS, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8380  -0.3328  -0.2159  -0.1679   3.4879

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.097770   0.168327 -12.462  < 2e-16 ***
trigger      -1.013414   0.059801 -16.946  < 2e-16 ***
POSADJ       -2.522421   0.374524  -6.735 1.64e-11 ***
POSADV       -1.520623   0.221814  -6.855 7.11e-12 ***
POSADVPRO     0.290763   0.767028   0.379 0.704630
POSANUM     -11.468297 535.411195  -0.021 0.982911
POSAPRO      -0.476598   0.307355  -1.551 0.120987
POSCNJ       -2.677258   0.308176  -8.687  < 2e-16 ***
POSCONJ      -0.636167   0.193254  -3.292 0.000995 ***
POSIMIT     -11.256841 159.777042  -0.070 0.943833
POSINTJ      -1.209730   0.603266  -2.005 0.044931 *
POSINTRJ    -11.419283 137.854950  -0.083 0.933982
POSN         -0.761782   0.168083  -4.532 5.84e-06 ***
POSNPRO      -0.738395   0.531599  -1.389 0.164830
POSNUM        0.562971   0.189674   2.968 0.002996 **
POSPARENTH    0.979778   0.288075   3.401 0.000671 ***
POSPART      -1.366386   0.218055  -6.266 3.70e-10 ***
POSPOST      -1.862565   0.309625  -6.016 1.79e-09 ***
POSPR        -0.008138   0.200163  -0.041 0.967571
POSPRAEDIC    1.231856   0.684428   1.800 0.071887 .
POSPRO       -2.969338   0.255873 -11.605  < 2e-16 ***
POSUNK       -0.768057   0.173638  -4.423 9.72e-06 ***
POSV         -1.143372   0.172869  -6.614 3.74e-11 ***
POSnorm      15.663836 535.411195   0.029 0.976661
---

    Null deviance: 18818  on 65498  degrees of freedom
Residual deviance: 17884  on 65475  degrees of freedom
AIC: 17932
```

Conclusion

- Trigger words facilitate code-switching

- Some parts of speech are more likely to be switched than others


Questions?