AKADEMIE DER
WISSENSCHAFTEN
IN HAMBURG

Individual code-switching strategies in language shift
The case of Nanai and Ulcha

Natalia Stoynova (stoynova@yandex.ru)
University of Hamburg

Moscow, IRL 09.02.2026

AKADEMIE DER WISSENSCHAFTEN IN HAMBURG

Introduction: Code-switching (CS) in an endangered language

→ An artificial situation when a linguist asks speakers to tell something in their native language, which is no longer used actively

- a lot of fragments in the dominant language
- a specific mode of code-switching: seems to differ in structural properties from fully spontaneous CS
- less studied than spontaneous CS
- At the same time, small corpora of endangered minority languages provide a lot of data of this type

Introduction: Inter-speaker variation in code-switching strategies

→ Inter-speaker variation of any kind is very typical of small speech communities in the situation of language shift

→ cf., e.g., Dorian 2010

→ Quantitative studies on inter-speaker variation in code-switching are rare

→ cf., however, Si & Ellisson 2023 on Hindi–English

→ In this talk:

- CS in oral texts in **Nanai and Ulcha** (Tungusic, endangered) with fragments in Russian (official language of the area)

Introduction: Aims of the study

To assess quantitatively **inter-speaker variation** in code-switching strategies (**structural types of CS**) used in texts in an endangered language collected from the last speakers

→ To reveal **clusters of speakers**

- Which speakers use the same CS strategies?
- How great is the variation?

→ To reveal **clusters of CS strategies**

- Which CS strategies determine inter-speaker variation?
- Which ones do it similarly?
- Which ones are stable across speakers?

→ (To explain the attested clusters)

Outline

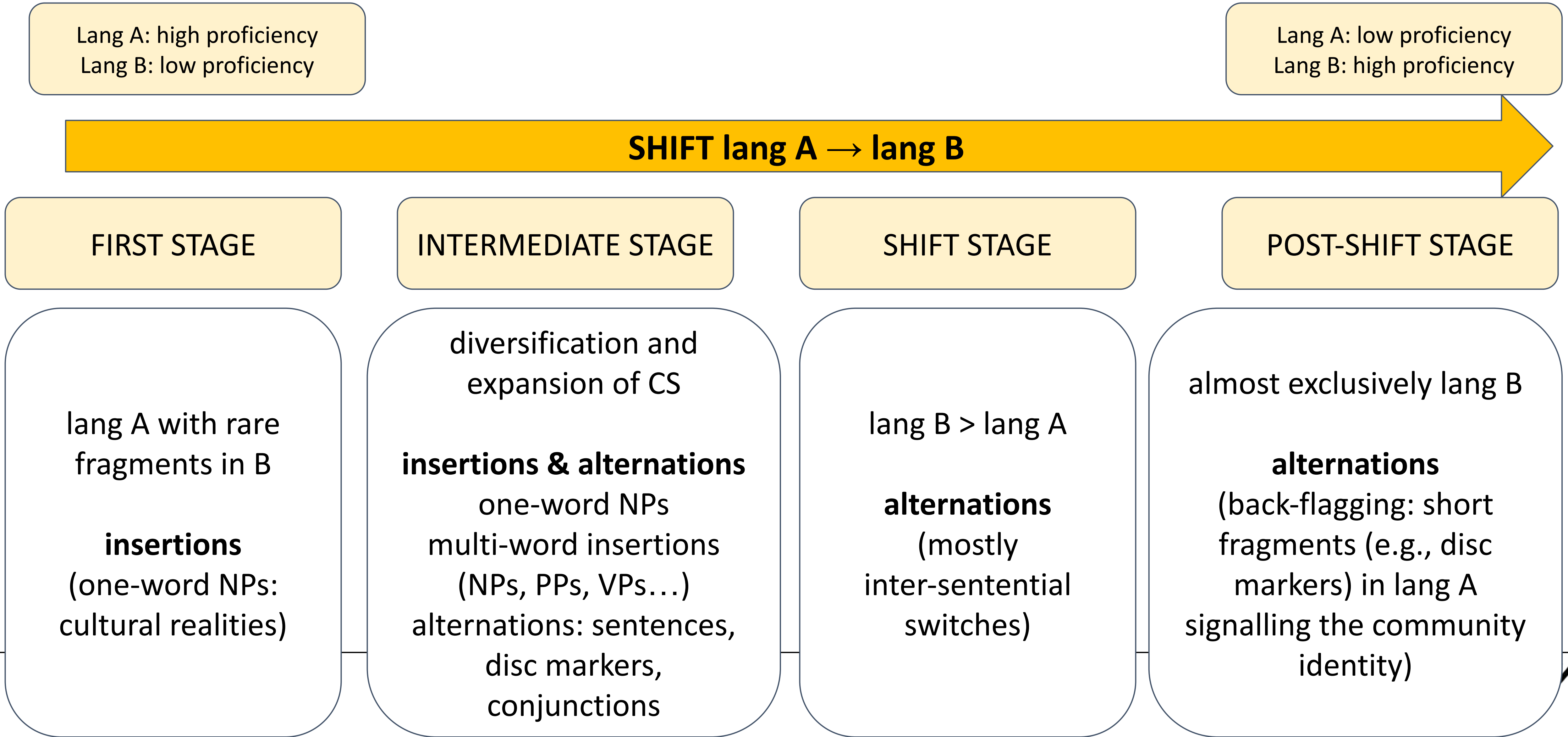
- Code-switching in language shift
- Data: Nanai and Ulcha texts with Russian fragments
- Annotation of code-switching
- Quantitative analysis
- Results
- Conclusions and discussion

Code-switching in language shift

Strategies of code-switching : Muysken 2000

INSERTION	**ALTERNATION**	**CONGRUENT LEXICALIZATION**
fragments of lang B are integrated (*inserted*) into the structure of lang A	well-formed separate fragments in lang A and lang B follow each other (*alternate*)	fragments in lang A and fragments in lang B fill the structure shared by A and B
morphosyntactically integrated constituents: NPs, PPs, Adj-s…	morphosyntactically non-integrated constituents: e.g., disc markers, sentences, non-constituents	*attested in CS between closely-related languages* no clear borders between A and B shared structure of A~B see Deuchar et al. 2007 on idenifying these types of CS
asymmetry and clear borders between lang A (**matrix**) and lang B (embedded)	symmetry and clear borders between lang A and lang B	

Code-switching in language shift (after Aalberse et al. 2019: 67–86)



Lang A: high proficiency
Lang B: low proficiency

Lang A: low proficiency
Lang B: high proficiency

SHIFT lang A → lang B

FIRST STAGE

lang A with rare fragments in B

insertions
(one-word NPs: cultural realities)

INTERMEDIATE STAGE

diversification and expansion of CS

insertions & alternations
one-word NPs
multi-word insertions (NPs, PPs, VPs…)
alternations: sentences, disc markers, conjunctions

SHIFT STAGE

lang B > lang A

alternations
(mostly inter-sentential switches)

POST-SHIFT STAGE

almost exclusively lang B

alternations
(back-flagging: short fragments (e.g., disc markers) in lang A signalling the community identity)

Code-switching in language shift

In this talk:

- What is observed at the 'shift' and 'post-shift' stages,
- **when a speaker is instructed/ consciously tries to "speak their language and not the dominant one"**?

What is known from previous research:

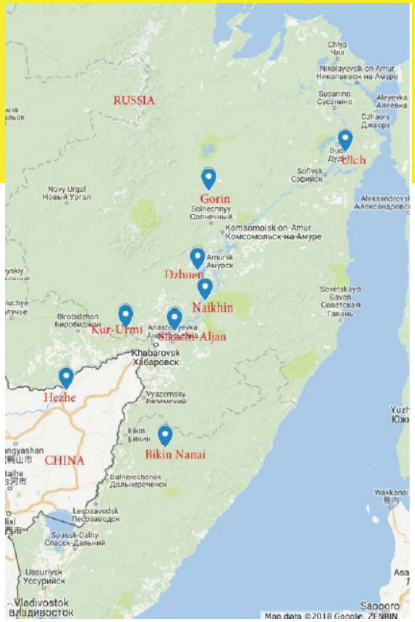
breaking borders between lang A and lang B

- non-standard structural patterns similar to **congruent lexicalization** (see Lipski 2014)
- "embedded language islands": non-standard "insertions" with lang B structure/inflection
- non-constituents
- no clear main/"matrix" language (see Myers-Scotton 1992; 2002)

see also on languages of Siberia, e.g., Grenoble 2010 (Evenki–Russian CS)

Data: Nanai and Ulcha texts with Russian fragments

Nanai and Ulcha



Map data ©2018 Google, ZENRIN

Tungusic: two closely-related sisters (Nanaic group)

The Amur region (Khabarovsk Krai, Russia)

Highly endangered

- a progressing shift to Russian (the official language of the region)
- all speakers also speak Russian, most of them use it more actively than Nanai/Ulcha
- no transmission to children, all speakers are of older generations

Nanai (Amur dialects)

- 1347 speakers, 11 % of the ethnic group (Census 2010)

see Gerasimova (2002); Kalinina & Oskolskaya (2016)

Ulcha

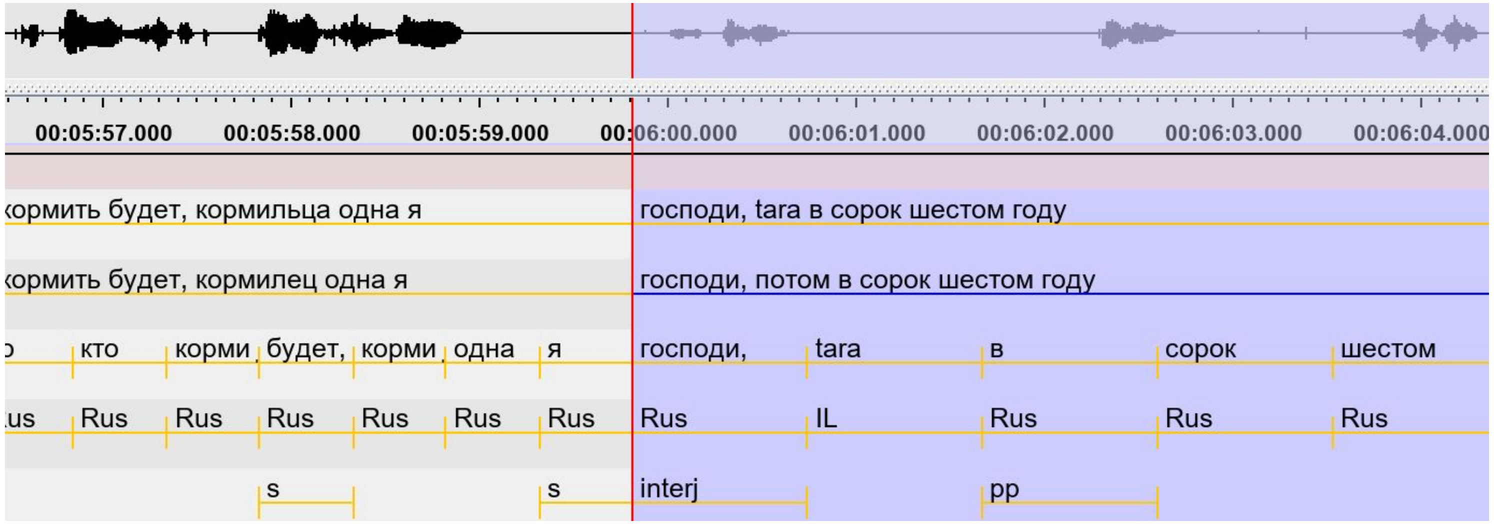
- 154 speakers, 6 % of the ethnic group (Census 2010)

see Gerasimova (2002); Sumbatova & Gusev (2016)

Data

Texts in Nanai and Ulcha recorded in the field (with Sofia Oskolskaya)

- transcribed and translated into Russian
- partly glossed
- Russian fragments (CS): annotated manually for size and morphosyntactic type (see Dyachkov et al. 2020 on the annotation)



00:05:57.000 00:05:58.000 00:05:59.000 00:06:00.000 00:06:01.000 00:06:02.000 00:06:03.000 00:06:04.000

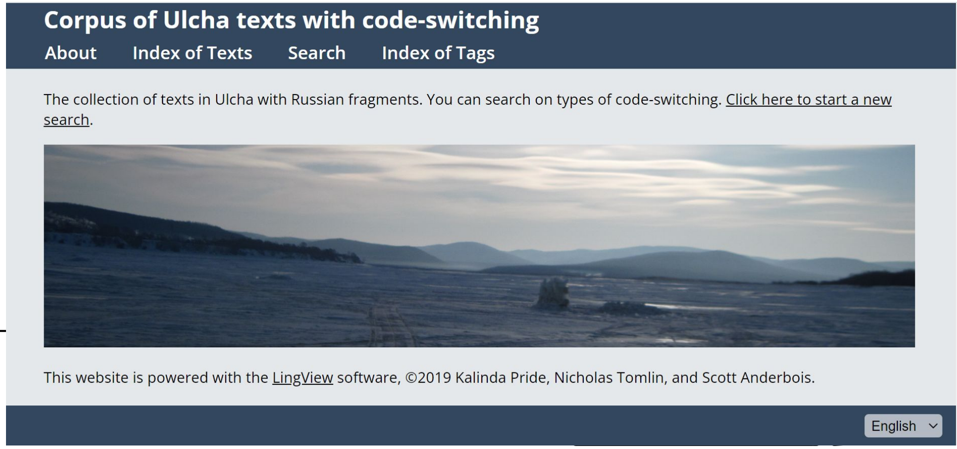
кормить будет, кормильца одна я | господи, tara в сорок шестом году

кормить будет, кормилец одна я | господи, потом в сорок шестом году

о | кто | корми | будет, | корми | одна | я | господи, | tara | в | сорок | шестом

Rus | Rus | Rus | Rus | Rus | Rus | Rus | Rus | IL | Rus | Rus | Rus

s | s | interj | pp



Corpus of Ulcha texts with code-switching

The collection of texts in Ulcha with Russian fragments. You can search on types of code-switching. Click here to start a new search.

This website is powered with the LingView software, ©2019 Kalinda Pride, Nicholas Tomlin, and Scott Anderbois.

Texts and speakers

Texts

- oral, spontaneous
- BUT: produced under a special instruction of the linguist (“to tell a story in the native language and not in Russian”)
- short narratives: life-stories etc.
- 108,817 tokens (ca. 25 hours)

Speakers

- of older generations: 1930-1961 years of birth (younger speakers do not produce texts)
- 53 speakers → 24 speakers (enough texts, enough sociolinguistic information)

Annotation of code-switching

Sample: 5,357 code-switches

→ Intrasentential code-switching (code-mixing) only

- Russian sentences and larger fragments were excluded

→ Code-switching in a broad sense: no differentiation between code-switching and borrowing

- one-word Russian fragments were included
- Russian words with Tungusic inflection were included

(1)	*Ti*	*tatuč-i*	***awgust***	*beː=tani*	***rybač**-i*	*bi-či-ti*
	that	learn-PTCP.PRS	Avgust	month=COORD	fish-PRS	be-PST-3PL

‘The school-children used to fish in August’ (oax, Ulcha)

Sample: 5,357 code-switches

- old established phonetically adapted loanwords were excluded

Ulc. *gumaska* < Rus. *bumažka* ‘rouble’
Ulc. *pilisi-* < Rus. *pljasa-tj* ‘dance’

- Russian proper names were excluded

Switched constituents

15 tags were used in the CS-annotation for switched constituents:

- ADJ, ADV, CONJ, DISC, INTERJ, NP, NUMP, PP, …

→ The most frequent types were included in this study.

- **DISC (disc_one & disc_multi)**

ALTERNATION

(1) *Mi ənulukəi, navernoe*

‘I guess, I'm sick!’ (jutsg, Nanai)

(2) *Cadu naj* vsʼo rawno *ǯobo-j*

‘People still work there’ (lkb, Nanai)

Switched constituents

→ The most frequent types were included in this study.

- **CONJ (conj_coord & conj_subord)**

ALTERNATION?

(1) *Gučkuli ili gučkuli biəsi?*

‘Is (he) good-looking or is not (he) good-looking?’ (itg, Nanai)

(2) *Potomu što piktəguj baariduji*…

ALTERNATION???

‘Because when one gives birth to a child…’ (itg, Nanai)

Switched constituents

- **NPs: 3 sub-types**

NP_one

(1) *Babuška wəndi bičin*

‘(My) grandmother used to say…’ (oax, Ulcha) – **one-word**

INSERTION

NP_multi

(2) *Ca=tani, mylo xozjajstvennoe ǯapaxa*

‘And he took this thing, laundry soap.’ (itg, Nanai) – **multi-word**

Switched constituents

NON-STANDARD
problematic, e.g., for 4-M model by Myers-Scotton
(Myers-Scotton 2004; Myers-Scotton & Jake 2009; 2017)

NP_morph_rus

(3a)	*Kopʼjom*,	*xaj*,	*waː-ri*	*bi-či-ti*
	spear.INS	what	kill-PRS	be-PST-3PL

‘They killed (a bear), so, with a spear’ (aid, Ulcha) – **with Russian inflection** (INS is expected in Ulcha)

Switched constituents

NON-STANDARD
problematic, e.g., for 4-M model by Myers-Scotton
(Myers-Scotton 2004; Myers-Scotton & Jake 2009; 2017)

NP_morph_rus

(3b)

Xaj-wa,	*baqam,*	
what-ACC	*find-CVB.SIM.SG*	
dekretnogo	*buː-rəs*	*bi-či-n=gun.*
*maternity.leave.**GEN.SG***	*give-NEG.PRS*	*be-PST-3SG=COMM*

‘After giving birth, one did not give us maternity leave’(oax, Ulcha) – **with Russian inflection** (ACC is expected in Ulcha)

(3c)

Vakansij	*kəwə*
position.**GEN.PL**	NEG.EX

‘There are no working positions. ’ (aid, Ulcha) – **with Russian inflection** (NOM is expected in Ulcha)

Word-internal switches: With Tungusic inflection

- **MORPH (word-internal)**

INSERTION

(1) *trjohlitrovaja banka-sal-či təučū-ri-ni=go*
three.liter jar-PL-LAT load-PRS-3SG=PTCL
‘One puts it to three-liter jars’ (rchk, Nanai)

(2) *a sin deda-ŋgu-s=gdəli…*
and your grandfather-ALIEN-2SG=EMPH
‘And your grandfather…’ (lpd, Ulcha)

(3) *pečem-bə-ni žari-la-go-o-ri*
liver-ACC-3SG grill-VBLZ-REP-IMPS-PRS
‘… One grills its liver’. (rchk, Nanai)

Non-constituent switches

- **Non-constituent switches**

NON-STANDARD (CL?)

(1)	*Mimbə*	*baqa-xa-n*	*ona*	*v*	*senjax*
	1SG.ACC	find-PST-3SG	she	in	porch.PL.LOC

‘She gave birth to me in the porch.’ (mkd, Ulcha) – **nonconst_integr**

(2)	*i*	*vot*	*i*	*siksə=dələ*	*naː-t*	*dəŋs-i*
	and	so	and	evening-ADVZ.LIM	3-3PL	work-PRS

‘And so and they work until late evening’(gip, Ulcha) – **nonconst_other**

ALTERNATION

Sentences with Matrix Language Russian

NON-STANDARD
(no consistent matrix language within the text)

- **v_rus** (≈Russian finite verb)

(1)	*Babka=ŋgu-s*	*sin-ti*	*ničego*	*ne*	*peredala?*
	grandmother-ALIEN-2SG	you.SG-LAT	nothing	NEG	PREF.give.PST.SG.F

‘Did not your grandmother transmit you anything (of her shaman skills)?’ (epv, Ulcha)

No further annotation for such sentences

see Bullock et al. (2018) for a discussion on different approaches to identifying the main (matrix, dominant) language of a sentence / text / corpus with CS

Individual features of speakers

→ Standard sociolinguistic parameters

- language (Nanai vs. Ulcha)
- year of birth
- level of education

→ Parameters ~ fluency in Nanai/Ulcha

- speech rate
- frequency of placeholders (*xaj, xajwa, eto*)

Individual features of speakers

→ Parameter ~ fluency in Russian – for 10 speakers only

- a morphosyntactic index showing to which degree the speaker's Russian differs from Standard monolingual Russian
- (normalized N of morphosyntactic deviations from Standard Russian attested in their Russian speech, based on our corpus Khomchenkova et al. 2019: http://web-corpora.net/ruscontact/corpus.html)

Quantitative analysis

WORK-IN-PROGRESS

Analysis

→ **Principal Component Analysis** (PCA)

- clustering structural types of code-switching (variables)
- variables are decomposed into “principal components” (dimensions) describing the variation between individuals (speakers) in the best way

↓

→ **Hierarchical clustering on principal components** (HCPC)

- clustering speakers (individuals)
- individuals are clustered in the multi-dimensional space of these principal components

cf. Husson et al. 2010; Abdi & Williams 2010; Levshina 2015: 353–361

Analysis

Active variables (in the analysis) – 9

- morph, disc, conj, np_multi, np_one, np_morph_rus, nonconst_other, nonconst_integr, v_rus

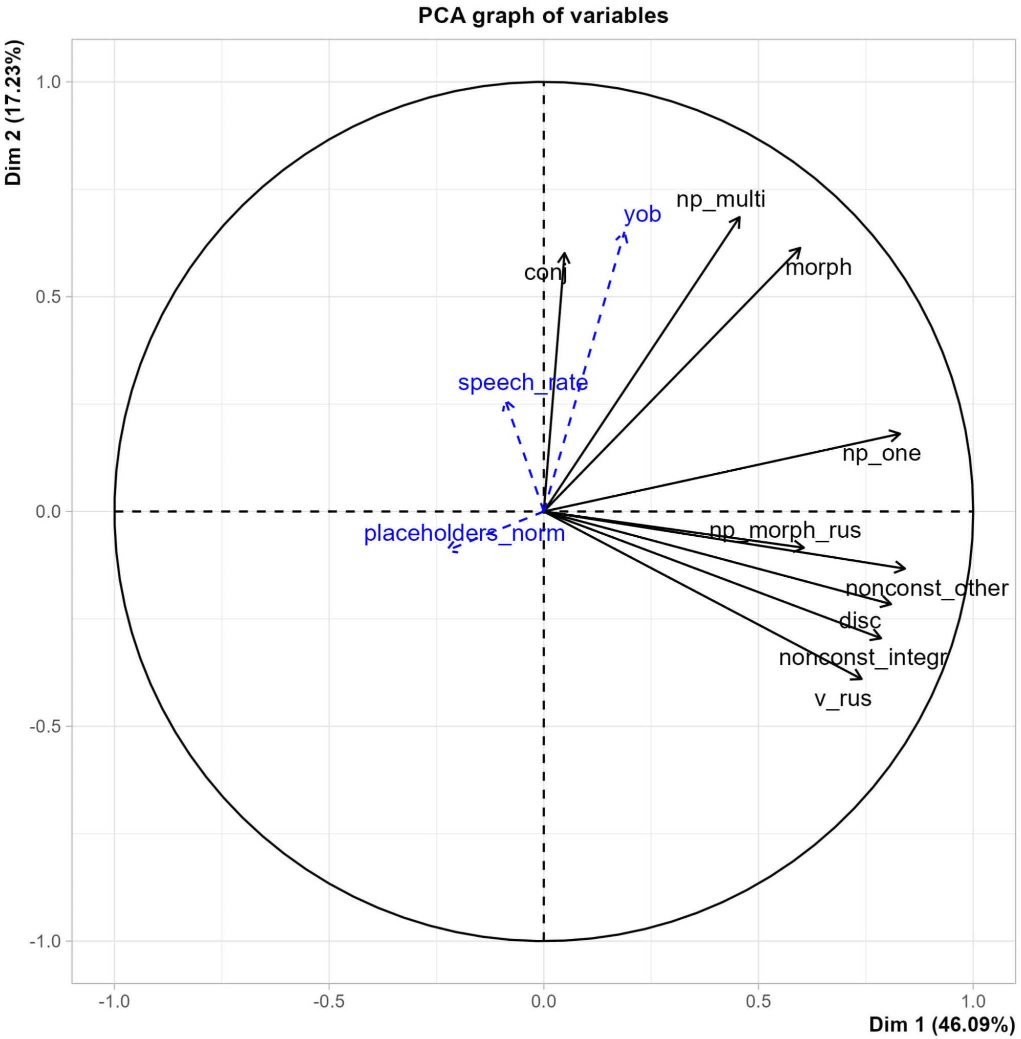
Supplementary variables (to see correlations)

- year of birth, speech_rate, placeholders, (rus_index)
- qualitative: language, education

N of switches of each type → normalized per 1,000 clauses

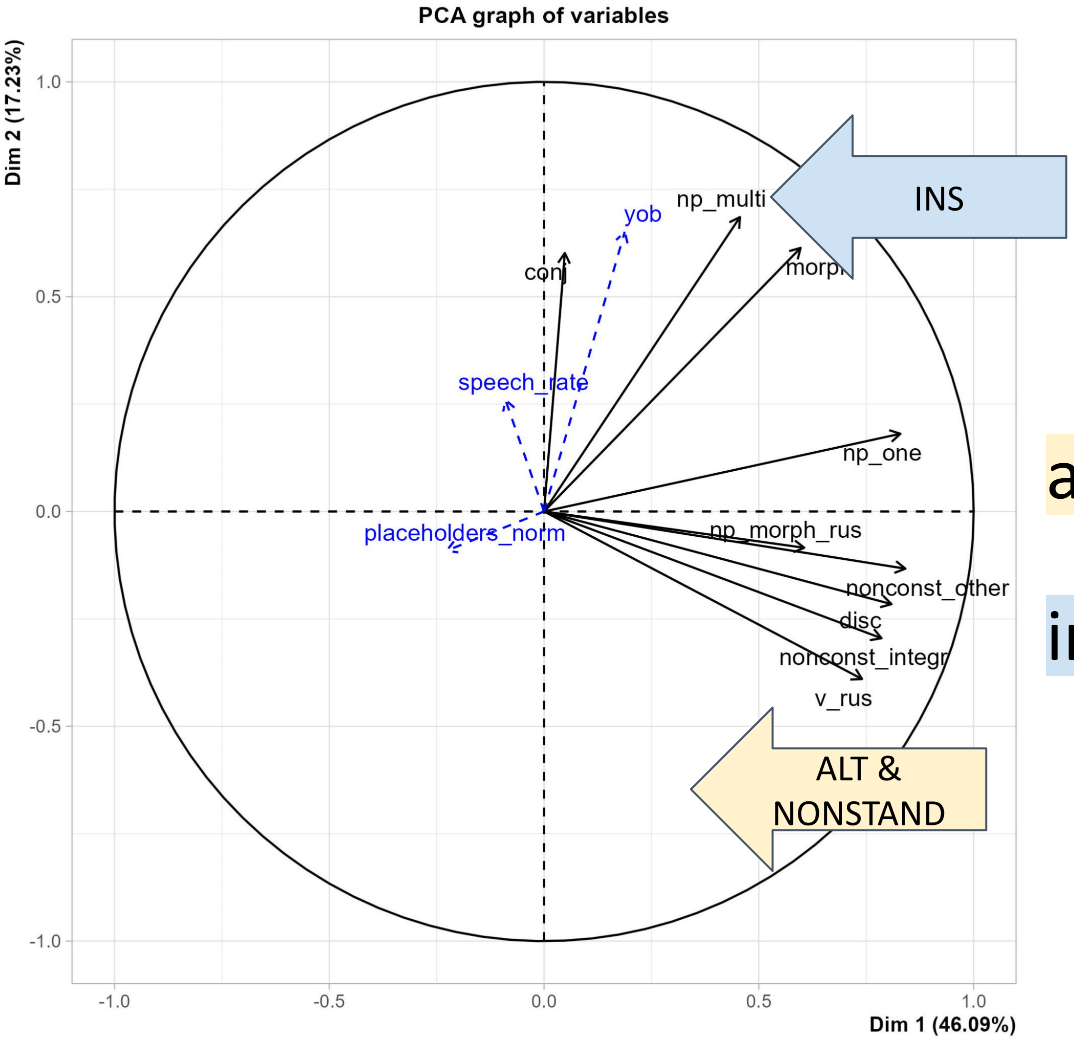
Results

Types of code-switching



→ **All describe the variation across speakers relatively uniformly = vary similarly across speakers**

Types of code-switching



→ **All describe the variation across speakers relatively uniformly = vary similarly across speakers**

→ <u>Variables behaving in a similar way</u>

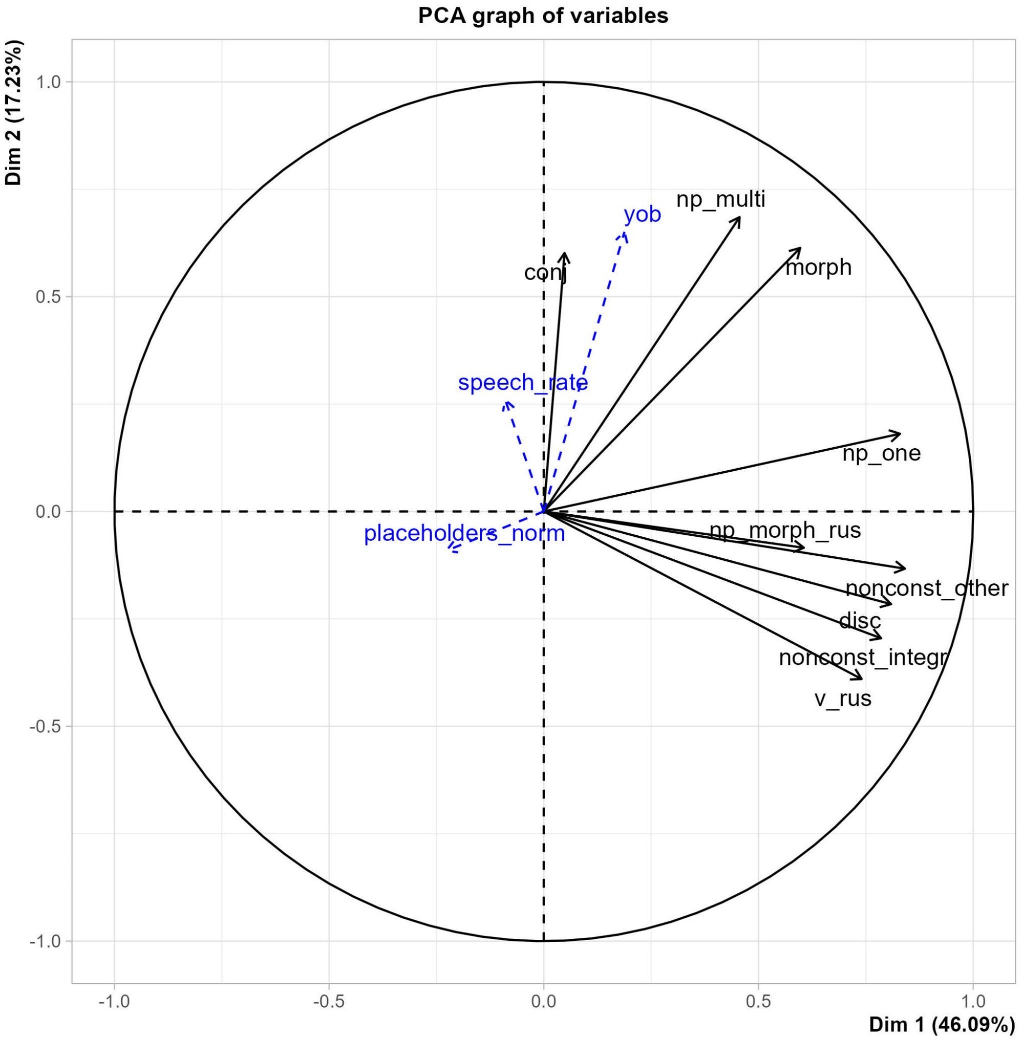
alternation & nonstandard

- nonconst, disc, v_rus, np_morph_rus

insertion

- np, morph, *conj*

Types of code-switching



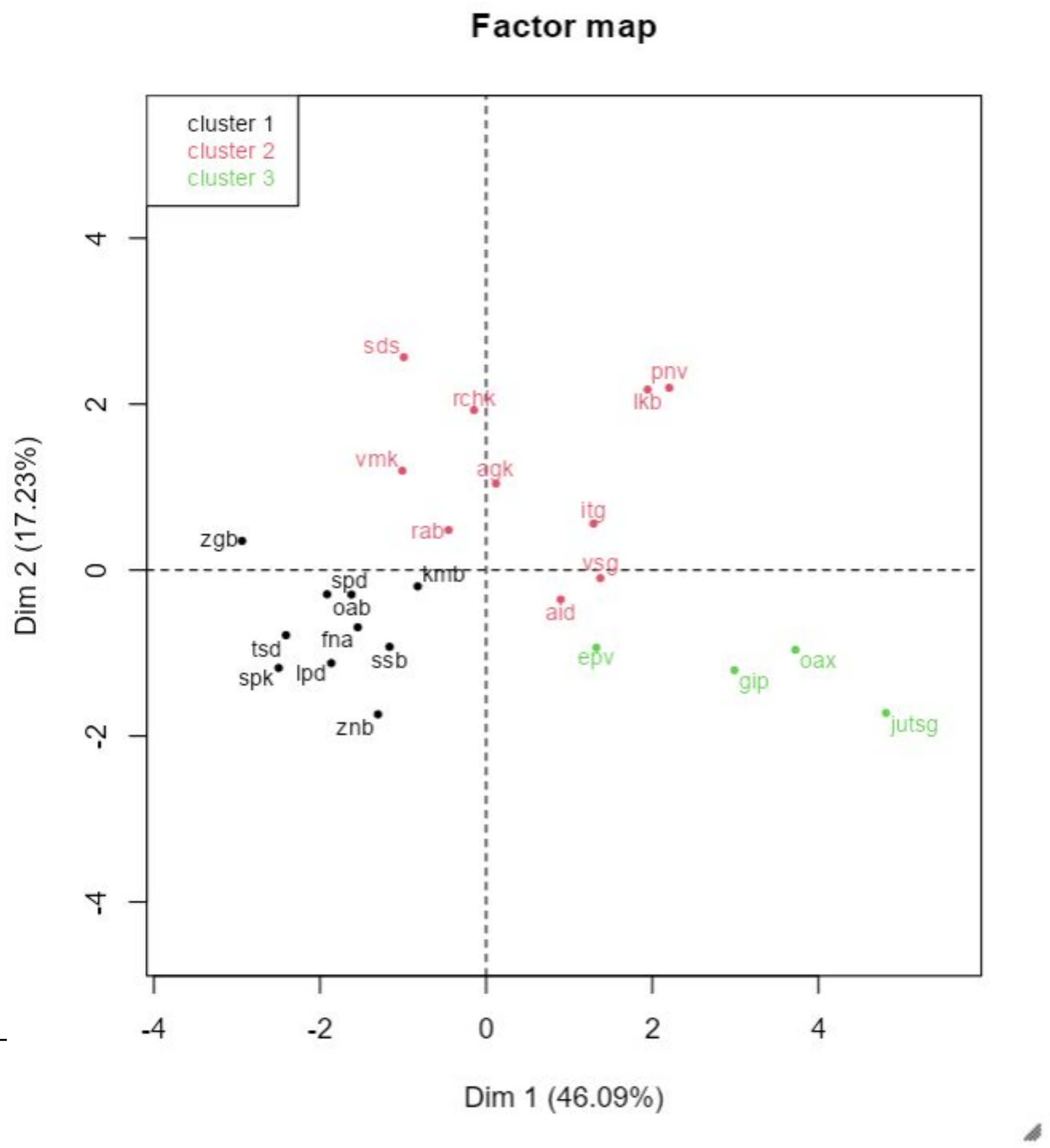
→ <u>Variables with the lowest contribution</u> = stable across speakers

- *np_morph_rus*, conj

→ <u>Variables with the highest contribution</u> = the most varying across speakers

- nonconst_other, np_one, morph

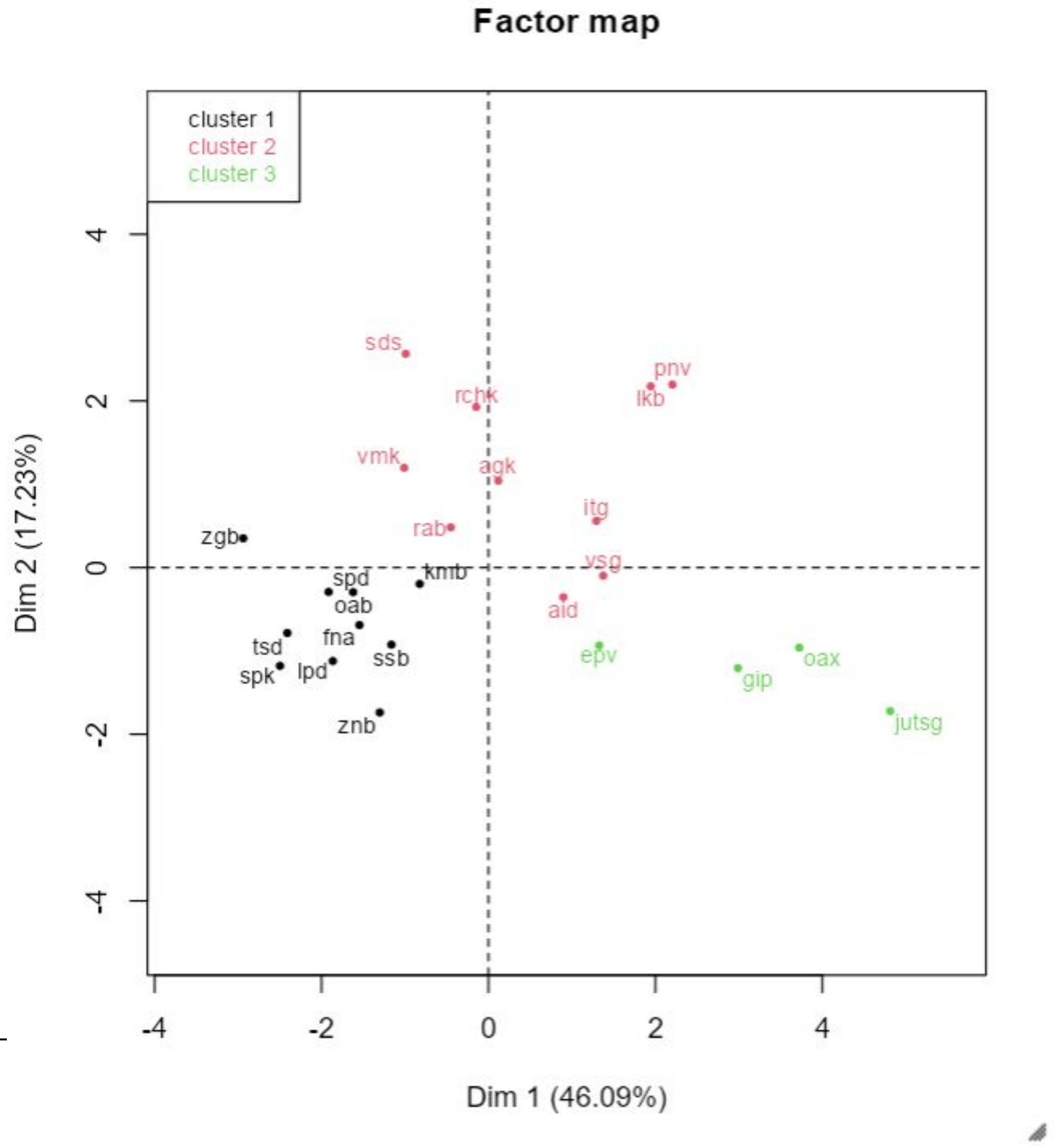
Clusters of speakers



→ **Cluster 1** “Non-switchers”

- all types of CS (except for conj and np_morph_rus) – significantly low
- correlates with age: older speakers

Clusters of speakers

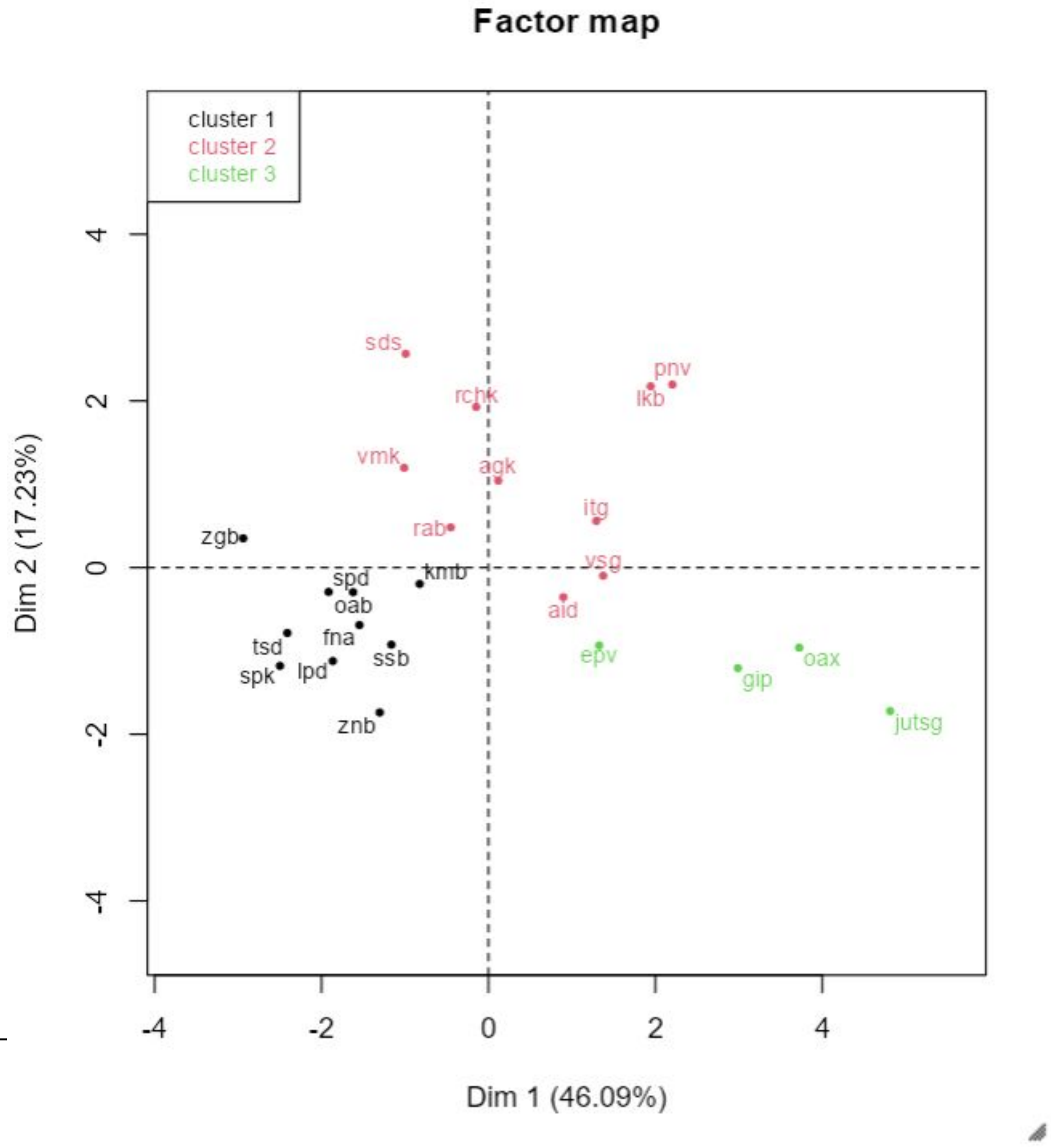


→ **Cluster 2**

“Inserters”

- np_multi, morph, conj – significantly high
- correlates with age: mostly younger

Clusters of speakers

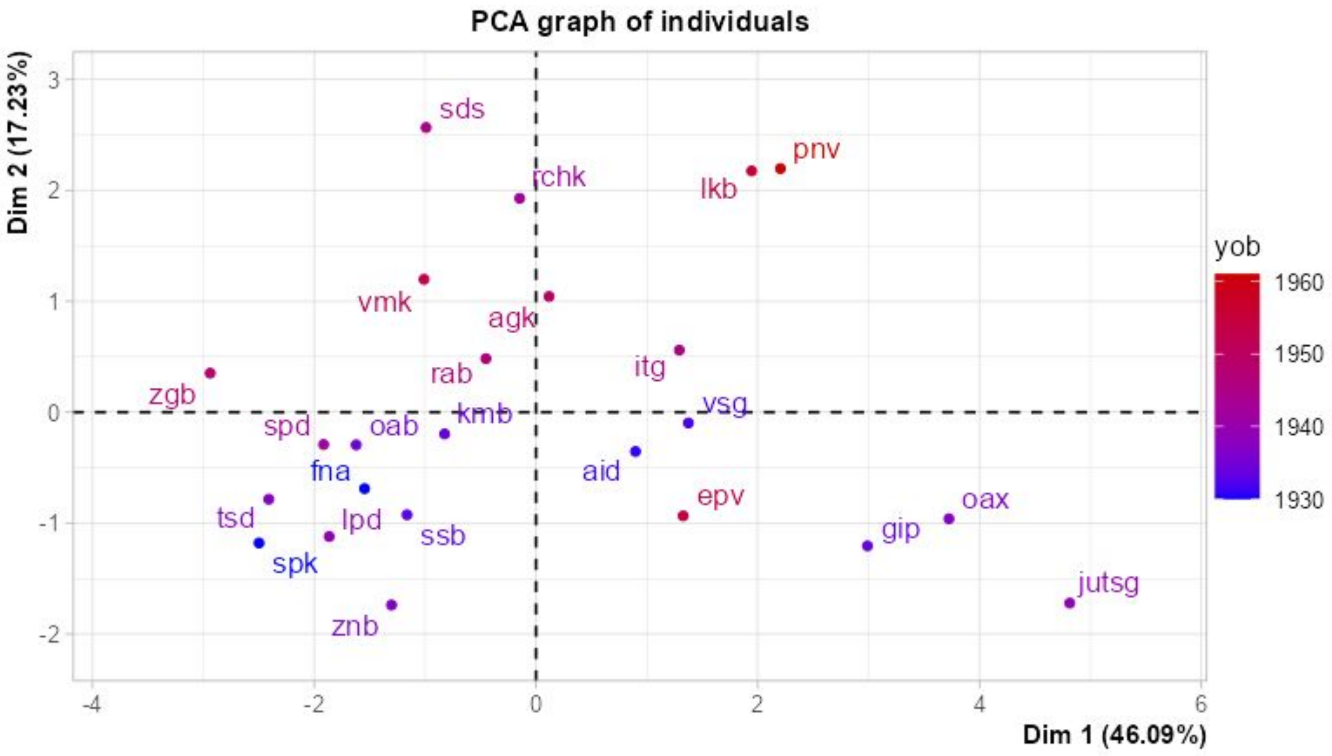


“Non-standard switchers”

→ Cluster 3

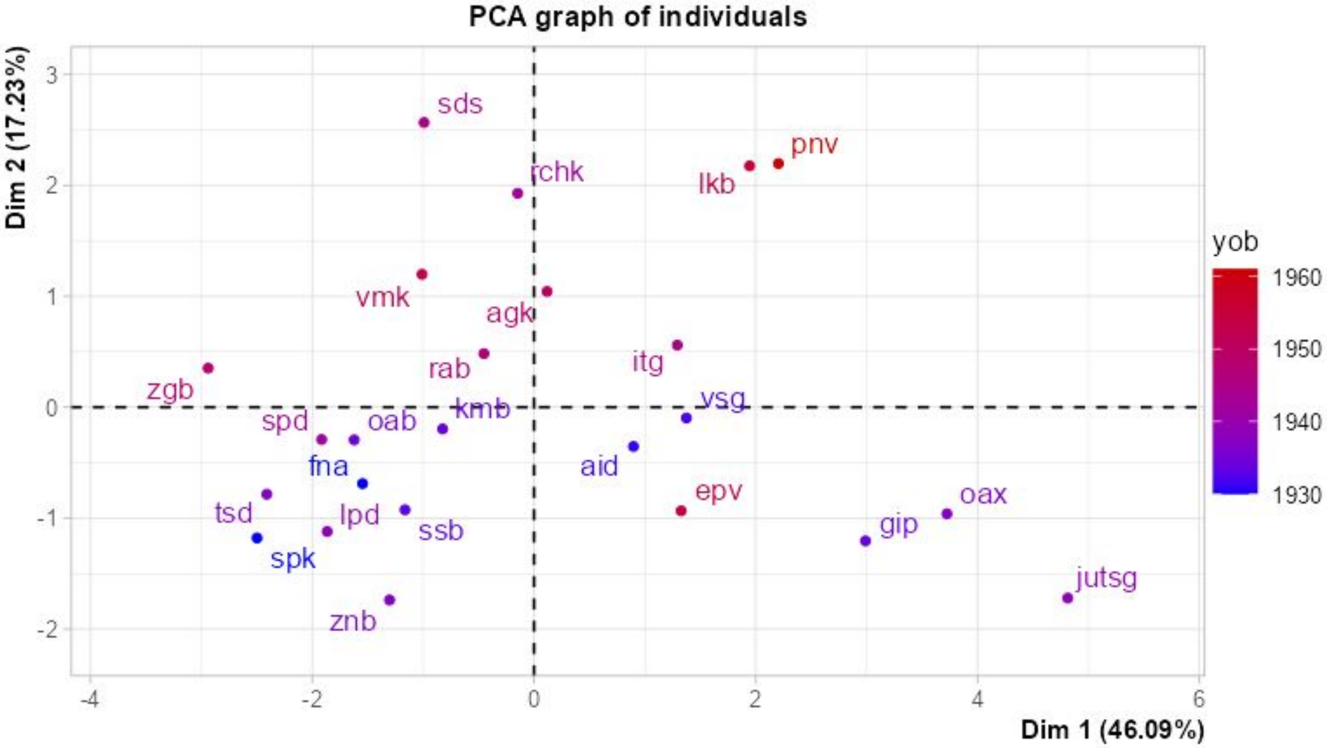
v_rus, disc, nonconst, *np_one*

Clusters of speakers



→ Correlation with year of birth

Clusters of speakers



→ No evident correlations with

- level of education
- speech rate
- frequency of placeholders
- morphosyntactic deviations from standard Russian

Conclusions & Discussion

Conclusions & Discussion

→ Different types of code-switching are do not differ a lot in describing inter-speaker variation

- however: alternations & non-standard switches vs. insertions

- the annotation is too rough to capture the most interesting features
  > more elaborated annotation needed?

Conclusions & Discussion

→ 3 clusters of speakers

- correlates with year of birth only
- to search for other predictors? more accurate annotation?
- the most interesting cluster of non-standard switchers: no visible correlations

Conclusions & Discussion

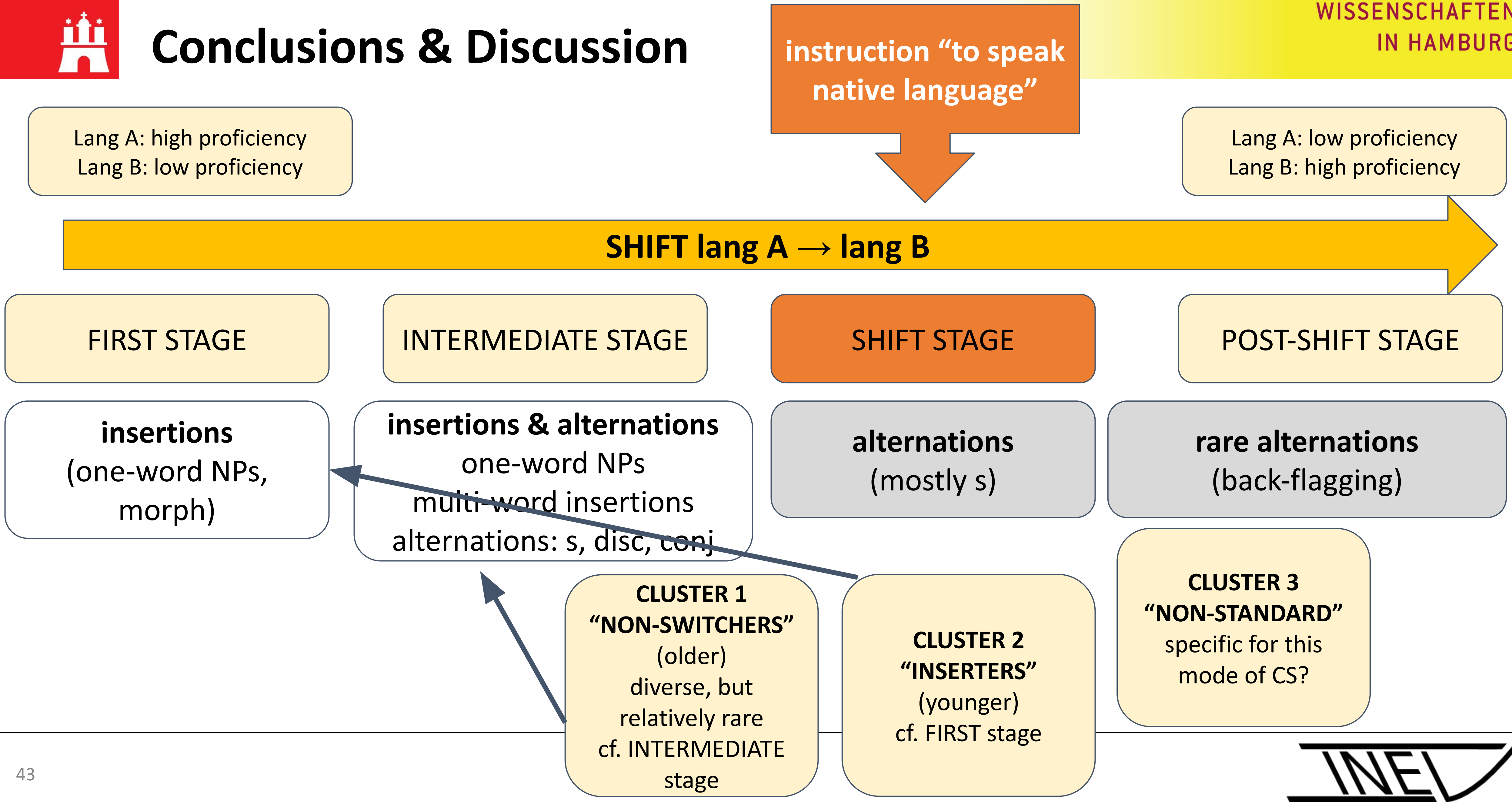
→ Lack of data

- rare types of code-switching – not included
- speakers with a small number of texts – not included

→ Lack of annotation

- N of switches per 1,000 clauses – a problematic measure
- e.g., >> disc = a speaker uses many switched discourse markers? a speaker uses many discourse markers?
- clauses/sentences – how to count?

Conclusions & Discussion



instruction “to speak native language”

Lang A: high proficiency
Lang B: low proficiency

Lang A: low proficiency
Lang B: high proficiency

SHIFT lang A → lang B

FIRST STAGE

INTERMEDIATE STAGE

SHIFT STAGE

POST-SHIFT STAGE

insertions
(one-word NPs, morph)

insertions & alternations
one-word NPs
multi-word insertions
alternations: s, disc, conj

alternations
(mostly s)

rare alternations
(back-flagging)

CLUSTER 1
“NON-SWITCHERS”
(older)
diverse, but relatively rare
cf. INTERMEDIATE stage

CLUSTER 2
“INSERTERS”
(younger)
cf. FIRST stage

CLUSTER 3
“NON-STANDARD”
specific for this mode of CS?

References

Aalberse, Suzanne, Ad Backus, and Pieter Muysken. 2019. Heritage Languages. A language contact Approach. Amsterdam/Philadelphia: John Benjamins.

Deuchar, M., Muysken, P., and S.L. Wang. 2007. Structured variation in codeswitching: towards an empirically based typology of bilingual speech patterns // International Journal of Bilingual Education and Bilingualism, 10. P. 298–340.

Dyachkov, V., Khomchenkova, I., Pleshak, P., and N. Stoynova. 2020. Annotating and exploring code-switching in four corpora of minority languages of Russia // Computational Linguistics and Intellectual Technologies, 20. P. 228–240. https://doi.org/10.28995/2075-7182-2020-19-228-240.

Gerasimova, A.N. 2002. Nanai and Ulch in Russia: a comparative characteristics of the sociolinguistic situation [Nanajskij I uljčskij jazyki v Rossii: sravniteljnaja harakteristika sociolingvističeskoj situacii] // Jazyki Korennyh narodov Sibiri [Languages of Indigenous Peoples of Siberia], 12.

Kalinina, E.Ju., and S.A. Oskolskaya. 2016. Nanai [Nanajskij jazyk] // Language and Society. An Encyclopaedia [Jazyk i obščestvo. Ènciklopedija]. Moscow: Azbukovnik. P. 293–296.

Lipski, J.M. 2014. Spanish-English code-switching among low-fluency bilinguals: Towards an expanded typology // Sociolinguistic studies 8(1), P. 23–55.

Muysken, P. 2000. Bilingual speech: A typology of code-mixing. Cambridge/New York: CUP.

Myers-Scotton, C. 1998. A way to dusty death: the Matrix Language turnover hypothesis. In: Grenoble LA, Whaley LJ, eds. Endangered Languages: Language Loss and Community Response. Cambridge University Press. P. 289-316.

Myers-Scotton, C, and J. Jake. 2009. A universal model of code-switching and bilingual language processing and production // Bullock, B. E, and A. J. Toribio (Eds.) The Cambridge handbook of linguistic code-switching. Cambridge: CUP. P. 206–357.

Myers-Scotton, C, and J. Jake. 2017. The 4-M model revisited: Codeswitching and morpheme election at the abstract level // International Journal of Bilingualism, 21(3). P. 340–366. https://doi.org/10.1177/1367006915626588

Myers-Scotton, C. 2002. Contact linguistics: Bilingual encounters and grammatical outcomes. Oxford/New York: OUP.

Myers-Scotton, C. 2004. Precision tuning of the Matrix Language Frame (MLF) model of codeswitching // Sociolinguistica, 18. P. 106–117.

Si, A., and T. M. Ellison. 2023. Inter-individual differences in Hindi–English code-switching: A quantitative approach // International Journal of Bilingualism, 27(3). https://doi.org/10.1177/13670069221085018

Sumbatova, N.R., and V.Ju. Gusev. 2016. The Ulcha language [Ul'čskij jazyk] // Language and Society. An Encyclopaedia [Jazyk i obščestvo. Ènciklopedija]. Moscow: Azbukovnik. P. 513–515.