Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2020"

Moscow, June 17–20, 2020

ANNOTATING AND EXPLORING CODE-SWITCHING IN FOUR CORPORA OF MINORITY LANGUAGES OF RUSSIA[1]

Dyachkov V. V. (hyppocentaurus@mail.ru)
Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

Khomchenkova I. A. (irina.khomchenkova@yandex.ru)
Russian Language Institute, Russian Academy of Sciences & Lomonosov Moscow State University, Moscow, Russia

Pleshak P. S. (polinapleshak@yandex.ru)
University of Maryland, College Park, USA

Stoynova N. M. (stoynova@yandex.ru)
Russian Language Institute, Russian Academy of Sciences & NRU Higher School of Economics, Moscow, Russia

This paper describes code-switching with Russian in four spoken corpora of minority languages of Russia: two Uralic ones (Hill Mari and Moksha) and two Tungusic ones (Nanai and Ulch). All narrators are bilinguals, fluent both in the indigenous language (IL) and in Russian; all the corpora are comparable in size and genres (small field collections of spontaneous oral texts, produced under the instruction to speak IL); the languages are comparable in structural (dis)similarity with Russian. The only difference concerns language dominance and the degree of language shift across the communities. The aim of the paper is to capture how the degree of language shift influences the strategy of code-switching attested in each of the corpora using a minimal additional annotation of code-switching. We added to each corpus a uniform annotation of code-switching of two types: first, a simple semi-automatic word-by-word language annotation (IL vs. Russian), second, a manual annotation of structural code-switching types (for smaller sub-corpora). We compared several macro-parameters of code-switching by applying some existing simple measures of code-switching to the data of annotation 1. Then we compared the rates of different structural types of code-switching, basing on annotation 2. The results of the study, on the one hand, verify and enhance the existing generalizations on how language shift influences code-switching

[1] Supported by RFBR grant № 18-312-00155.

strategies, on the other hand, they show that even a very simple annotation of code-switching integrated to an existing field records collection appears to be very informative in code-switching studies.

Keywords: corpus linguistics, quantitative linguistic studies, language shift, code switching, code-switching metrics, Uralic languages, Tungusic languages

DOI: 10.28995/2075-7182-2020-19-228-240

РАЗМЕТКА И ИССЛЕДОВАНИЕ ПЕРЕКЛЮЧЕНИЯ КОДОВ В ЧЕТЫРЕХ КОРПУСАХ МАЛЫХ ЯЗЫКОВ РОССИИ[2]

Дьячков В. В. (hyppocentaurus@mail.ru)
ИЯз РАН, Москва, Россия

Хомченкова И. А. (irina.khomchenkova@yandex.ru)
ИРЯ РАН & МГУ им. М. В. Ломоносова, Москва, Россия

Плешак П. С. (polinapleshak@yandex.ru)
Университет Мэриленда, Колледж Парк, США

Стойнова Н. М. (stoynova@yandex.ru)
ИРЯ РАН & НИУ ВШЭ, Москва, Россия

1. Introduction

The aim of this paper is to show how frequency and structural types of code-switching correlate with sociolinguistic parameters, using quantitative data of the spoken corpora of several minority languages of Russia.

The use of two or more languages within one conversation or even within one utterance is known as the phenomenon of code-switching (CS). In this project, we understand it quite broadly (cf. [Treffers-Daller 1991]; [Myers-Scotton 1992]). We take into consideration all types of inter-clausal switching (1), as well as intra-clausal switching of constituents of different length (*i* 'and', *ixnjuju familiju* 'their last name' in (2)) and nonce borrowings from Russian, which often bear morphological affixes of the main language of the text (*veləs'ipecə* [bike.R.IN] in (1), *sestra-ni* [sister.R-3SG] in (2)).

[2] При поддержке гранта РФФИ № 18-312-00155.

(1) *veləs'ipecə* *ar-n'ə-s'-t'*.
bike.R.IN run-IPFV-PST.3-PL
oj, *togda* *velosiped-ov* *ne* *bylo*
oh.R then.R bike-PL.GEN.R NEG.R be.PST.3SG.N.R
'We were riding the bicycles. Oh, there were no bicycles that time'. (baa, Moksha)

(2) *i* *tị* *sestra-nị* *tị* *aldač-ị* *bi-či-n*
and.R that sister.R-3SG so tell-PRS be-PST-3SG
ixnjuju *familiju*
their.SG.F.ACC.R last.name.SG.ACC.R
'And his sister mentioned their last name.' (aid, Ulch)

Four languages of Russia were chosen for the research: Moksha, Hill Mari (Uralic); Nanai, and Ulch (Tungusic). They belong to different language families, but their structural (dis)similarity with Russian, which is relevant for our study, is more or less equal. For this study, we used the corpora that had been created by larger teams (including the authors) in the field during documentation projects on the corresponding languages, and the aim of the narrator was to tell a story in the indigenous language (IL). The text collections are comparable in sizes and genres (see below). For all the narrators, IL is the first language, or it was acquired simultaneously with Russian. All the speakers are highly proficient in Russian. They are also proficient in IL enough to tell a spontaneous story. So, although the degree of speakers' proficiency differs a lot on the level of the whole speech communities (see below), it is comparable for our text samples.

The only crucial difference among our datasets is the current use of the languages, which can be interpreted in terms of language dominance and language shift. The Hill Mari speakers use IL in their everyday communication at least as frequently as Russian. The Ulch narrators use IL much more restrictively than Russian or do not use it at all. In the Moksha and Nanai samples the situation varies across speakers. On the community level, across the speakers of Hill Mari a stable balanced bilingualism takes place, while the Moksha speakers, the Nanais, and especially the Ulchas are undergoing a progressing language shift to the dominant Russian language [Koryakov & Kholodilova 2018]; [Kalinina & Oskolskaya 2016]; [Gerasimova 2002]; [Sumbatova & Gusev 2016]. This can be represented as a hierarchy of language shift, cf. (3); see also **Table 1**:

(3) (language shift) Ulch > Nanai > Moksha >> Hill Mari (no language shift)

Table 1. Sociolinguistic information

language	N of speakers [Census 2010]	% of the ethnic group
Ulch	154	6%
Nanai	1,347	11%
Moksha	2,025	43%
Hill Mari	23,062	98%

Basing mostly on generalizations made in [Benthalia & Davies 1992], [Backus 1996], and [Muysken 2000: 227–228; 247–248], we have the following expectations

on possible correlations between language dominance and inter-generation shift[3], on the one hand, and structural types of CS, on the other hand:

- inter-clausal switches are more frequent in balanced bilinguals, than in the situation of dominance asymmetry;
- word-internal switches are more frequent in the situation of dominance asymmetry, than in balanced bilinguals;
- in the situation of language shift, the number of more syntactically integrated constituents (insertions in terms of [Muysken 2000]) increases, while the number of less integrated constituents (alternations in terms of [Muysken 2000]) decreases;
- in the situation of language shift the number of non-constituent switches (which also belong to alternations in terms of [Muysken 2000]) decreases.

Our general hypothesis was that, according to hierarchy (3), the main difference in CS strategies would be between the Hill Mari and Ulch corpora, as they represent the opposite sociolinguistic situations, with Nanai and Moksha corpora in between, having an intermediate stage of the language shift. To reveal specific properties of CS that vary across the text collections under discussion, we annotated each word for the language (IL vs. Russian) and added a specific annotation of structural types of CS (**Section 2**). Then, for each corpus we conducted calculations, based on language annotation, using some existing metrics developed for corpus-based studies on CS (**Section 3**). After that, for a smaller part of the text collection, we conducted more precise calculations, based on our annotation of structural types of CS (**Section 4**). Finally, we compared the results of the calculations for our four corpora, checked how they match to the language shift hierarchy, and related the correlations, attested in our data, to the existing observations (**Section 5**).

2. Annotation of code-switching types

All the text collections were annotated in ELAN using the same set of tiers and labels. The annotation tiers are the following: LANG, which indicates the language of each word (token), CS_TYPE, which indicates the syntactic type of the switched fragment (all tags are aligned to words). The annotated collections are available at: http://web-corpora.net/tsakorpus_russian_nonst/CS.html.

The LANG tier contains two tags: IL—indigenous language and Rus—Russian (including Russian words with IL-affixes). This tier was created semi-automatically, based on the script: the main transcription in our corpora is in Latin, while the majority of Russian fragments are in Cyrillic. So, the tag Rus was first assigned automatically to all words[4] transcribed in Cyrillic, then some tags were changed or added manually.

[3] Note, however, that they discuss mostly inter-generation shift within local communities, while language shift (i.e. the loss of language within the whole language community) is much less studied from this point of view.

[4] The word was recognized simply as an item separated by space-bars. Intermediate cases, such as clitics, were treated according to the writing system, adopted in each particular corpus. This might create some discrepancies in our data, but they are rather minor.

The CS_TYPE tier contains syntactic tags, which were assigned manually to a smaller part of the text collection (see Table 2 for general information on the corpora). They are listed in Table 3.

Table 2. Corpora: sizes and types of annotation

	provided with LANG tags, texts (tokens)	provided with CS_TYPE tags, texts (tokens)	other features of the corpus
Nanai	167 (47,411)	52 (16,368)	synchronized with audio
Ulch	179 (47,509)	50 (11,334)	synchronized with audio
Moksha	53 (17,578)	53 (17,578)	glossed
Hill Mari	17 (15,895)	17 (15,895)	glossed

Table 3. CS-type tags

tag	description
adj(+)[5]	adjectival phrase
adv(+)	adverbial phrase
conj(+)	conjunction phrase
dep	dependent clause
disc(+)	discourse marker
ideoph(+)	ideophone
interj(+)	interjection
morph	Russian stem with IL-affixes
morph_p	Russian multi-word phrase marked with IL-affixes
np(+)	noun phrase
nump(+)	numeral phrase
pp(+)	prepositional phrase
pred(+)	predicative word
s	sentence
v_rus[6]	clause with Russian verb
voc(+)	vocative forms
vp(+)	verb phrase
other	other constituent types

[5] "+" stands for multi-word constituents. In this case, the tag is assigned to the head.

[6] The texts under consideration are positioned by narrators as texts in the corresponding indigenous language (IL) and not in Russian, and the total amount of Russian fragments is much smaller than those in the IL. However, a potential possibility to reveal the main language ("matrix", ML) and the secondary one ("embedded", EL) for each particular clause with intrasentential CS is a matter of theoretical discussion (cf. [Myers-Scotton 1993: 46–74]; [2002: 15–16; 58–69]; [Muysken 2000: 1–34]). Our technical solution is to mark Russian fragments as switched (i.e. to consider the IL as ML) in all mixed clauses, except for those with Russian finite verbs (which are much less numerous in the sample). The latter are treated separately and take a tag *v_rus* with no further annotation.

Russian fragments that do not form any syntactic constituent are marked with corresponding tags separately (*conj*, *conj*, *pp* and *adj*) in (4).

(4) [*no*] [*i*] [*do* *vojny*] [*molodaja*] *bi-či-ni=goa*
but.R and.R before.R war.GEN.R young.SG.F.R be-PST-3SG=PTCL
'But before the war she also was young.' (itg, Nanai)

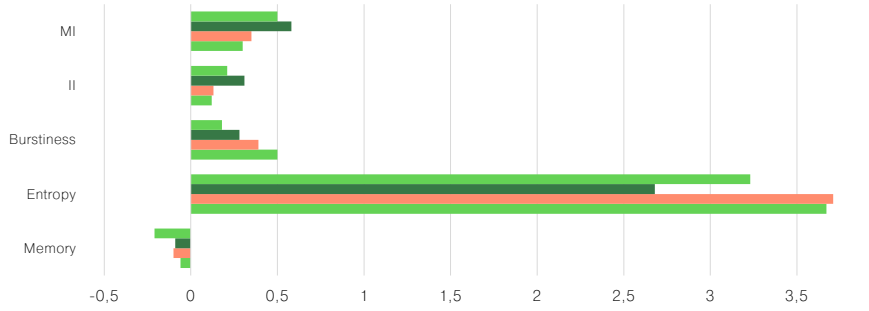
3. Metrics based on the word-by-word language annotation

The general information on CS, which allows us to compare the four text collections, was obtained from the word-by-word language annotation, available for the whole corpora (cf. **Section 2**). To characterize CS patterns, we used the existing metrics, proposed for corpus-based studies on CS and summarized in [Guzmán et al. 2016], [2017a,b]. Some of them are based on the ratio of L1-words and L2-words, some others are based rather on the ratio of L1-spans and L2-spans (where a L1-span is a word sequence in L1 bounded between L2-words). The general information on these metrics is given in **Table 4**.

Table 4. CS metrics based on word-by-word language annotation

Metric	Description	Formula	[from…; to…]	Reference
Multilingual index	measures how "bilingual" the text is: the (in)equality of the distribution between L1 and L2	$M\text{-}I = \frac{1-\sum p_j^2}{\sum p_j^2}$	[0; 1] [all words in L1; L1 and L2 in equal proportions]	[Barnett et al. 2000], [Gardner-Chloros et al. 2007], [Guzmán et al. 2016] [2017a]
Integration Index	measures how "bilingual" the text is: the probability of L1 vs. L2 within the text	$I\text{-}I = \frac{1}{n-1}\sum_{1\le i<j\le n} S(l_i, l_j)$	[0; 1] [L1, L1, L1…; L2, L1, L2,…]	[Gambäck & Das 2014, 2016]; [Guzmán et al. 2016], [2017a]
Burstiness	measures how (non)-random switches are: the regularity of switching spans	$Burstiness = \frac{(\sigma_\tau - m_\tau)}{(\sigma_\tau + m_\tau)}$	[−1; 1] [regular heartbeat-like switching; irregular switching]	[Goh & Barabási 2008]; [Guzmán et al. 2017a]
Language Span Entropy	measures how predictable language spans are: how many bits of information are needed to describe the distribution of language spans	$Span\ Entropy = -\sum_{i=1}^{M} p_i \log_2 p_i$	[0; log2(M), M = the number of possible span states] [all spans are of equal length; spans are of a different length]	[Guzman et al. 2017b]
Memory	measures how (non)-random switches are: whether the length of L1-span correlates with the length of the preceding L2-span	$Memory == \frac{1}{n_\tau - 1}\sum_{i=1}^{n_\tau - 1}\frac{(\tau_i - m_1)(\tau_{i+1} - m_2)}{\sigma_1 \sigma_2}$	[−1; 1] [short La spans are preceded by long L2 spans; short L1 spans are preceded by short L2 spans]	[Goh & Barabási 2008]; [Guzmán et al. 2017a]

Figure 1 contains the values of these measures for all our corpora. While calculating, initial and final Russian fragments were omitted.



	Memory	Entropy	Burstiness	II	MI
Hill Mari	-0,21	3,23	0,18	0,21	0,5
Moksha	-0,09	2,68	0,28	0,31	0,58
Nanai	-0,1	3,71	0,39	0,13	0,35
Ulch	-0,06	3,67	0,5	0,12	0,3

Fig. 1. Measures of CS: the data of the four corpora

The Multilingual Index (MI) is a word-count-based measure that quantifies the inequality of the distribution of language tags in a corpus. According to it, the Ulch corpus is the most monolingual, while the Moksha corpus is the most bilingual (Ulch < Nanai < Hill Mari < Moksha). Integration Index (II) is a metric that describes the probability of switching within a text. Languages with the same MI can have different number of switches (compare $[IL, IL, Rus, Rus]_1$ and $[IL, Rus, IL, Rus]_2$ with both MI=1, but II_1=0,(3) and II_2=1). The IIs of our four corpora correspond to their MIs and form the same hierarchy.

We also calculated metrics reflecting the distribution of CS across the corpus using language spans—the distance between switch points, i.e. the length of monolingual discourse. The Burstiness measures whether CS has a periodic character or occurs in bursts, i.e. how predictable switches are. All our corpora have unpredictable patterns of switching with the following hierarchy: Ulch > Nanai > Moksha > Hill Mari. The switching patterns in Ulch and Nanai are more unpredictable, while that in Hill Mari is the most predictable. In order to take into account the time ordering of the language spans, we calculated the Memory Index, which shows to which extent the length of language spans influences the length of following spans. The hierarchy of corpora is exactly the same as for the Burstiness. All language spans are rather unpredictable, but Hill Mari language spans are more negatively correlated, while Ulch language spans are more positively correlated. The Span entropy returns how many bits of information are needed to describe the distribution of the language spans. The hierarchy is a bit different: Nanai ≥ Ulch > Hill Mari > Moksha, so it does not correlate with the Burstiness and Memory Index, but rather similar to those for MI and II (although with the opposite direction).

Thus, two out of five measures, i.e. those operating with spans, give the results, more or less correlating with the language shift hierarchy in (3), while three others show different results, which, however, are all similar to each other.

4. Structural types of code-switching

For the manually annotated sub-corpora, we compared frequencies of different structural types of switched fragments and frequencies of fragments of different sizes. The frequency distribution for switched fragments' sizes (morpheme vs. one-word vs. multi-word vs. clause[7]) is given in **Figure 2**.

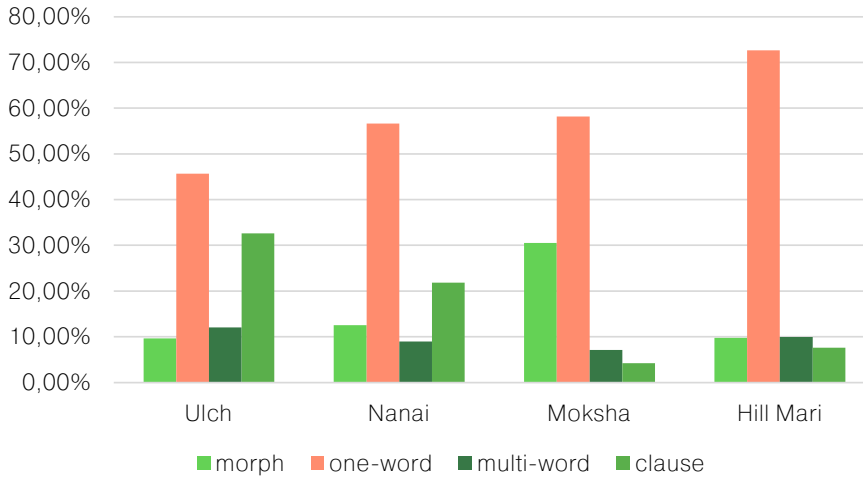


Fig. 2. Switched fragments' sizes

The rates of multi-word switches are comparable in all the collections and relatively low. The rates of other types of switches vary across the collections. Word-internal switches are much more frequent in Moksha than in all other languages. The percentage of clausal and one-word switches correlates with the language shift hierarchy (3). One-word switches form the most frequent type in all the samples, and their rate is lower in languages more affected by language shift (although the difference is quite modest). In contrast, clausal switches are much more frequent in languages more affected by language shift.

For switched constituents (excluding Russian stems with IL-affixes and Russian sentences), we calculated the frequency distribution of different syntactic types, see Figure 3. Only frequent types (> 1%) were included.

The most frequent types of switched constituents in all the corpora are *np*(+) (noun phrases), *disc*(+) (discourse markers), *conj*(+) (conjunctions) and *adj*(+) (adjectives). We expect the ratios to reflect the language shift hierarchy, introduced in (3). Across the frequent types, the correlation is attested for NPs and for discourse markers. Switched NPs are especially frequent in Ulch and the least frequent in Hill Mari; in opposite, discourse markers are the most frequent in Hill Mari and the least frequent in Ulch.

[7] Multi-clause switched fragments were not treated separately. Each of them was counted as several independent switched clauses. The same is true for multi-word switched fragments that do not form a syntactic constituent: they were counted as several independent constituents.

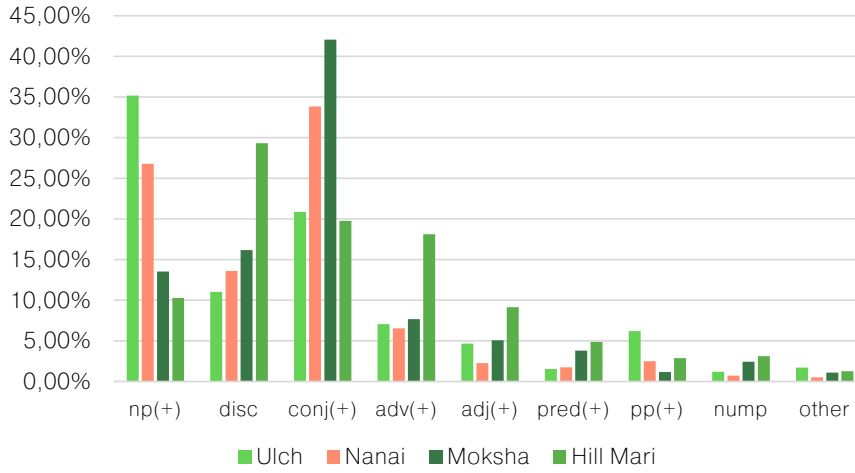


Fig. 3. Syntactic types of intra-sentential CS

This effect seems to be connected with the degree of syntactic integration. Discourse markers are elements which are not integrated into the syntactic structure, are uninflected forms and do not bear any overt markers of syntactic dependency, unlike NPs that are highly integrated into the clausal structure. We calculated the total ratio for non-integrated and integrated elements where we treated constituents of types *np*(+), *pp*(+), *nump*(+) as integrated and *disc*(+), *pred*(+), *interj*, *adv*(+), *voc*(+), and *adj*(+) as non-integrated[8]. For this total ratio, the correlation with the sociolinguistic hierarchy appears to be even clearer, see Figure 4.

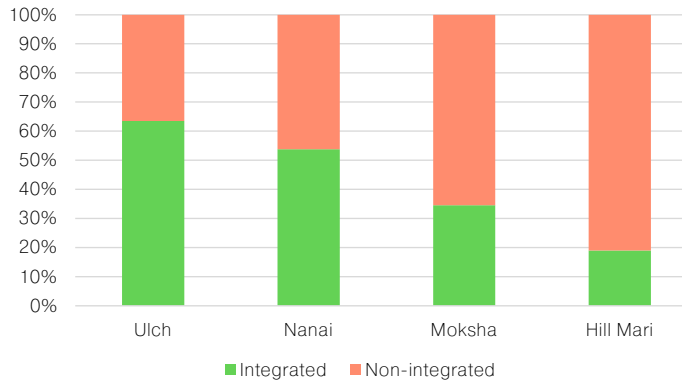


Fig. 4. Ratio of integrated and non-integrated elements

Integrated switched elements are typical of the language shift situation (Ulch), while non-integrated ones are typical of the situation of stable balanced bilingualism (Hill Mari).

[8] Conjunctions, which do not form part either of the two types of elements, were excluded.

5. Results and discussion

The crucial difference between the text collections under discussion concerns language dominance, i.e. the degree of language shift attested in the community. The hypothesis was that structural differences in CS would follow a hierarchy of languages that reflects the difference in their sociolinguistic status:

(5) (=3) Ulch > Nanai > Moksha >> Hill Mari

We applied to our data several simple measures of CS, based on the distribution of Russian words (Multilingual Index, Integration Index) and word-sequences (spans) across the text (Burstiness, Entropy, and Memory). Then, we checked whether they are interpretable in terms of language shift hierarchy. The measures based on switched spans appear to correlate with this hierarchy better, than those based on switched words.

We also checked some more fine-grained parameters of CS: the rate of clausal switches, the rate of word-internal switches, and the rate of different syntactic types of switched constituents. Not all of the attested asymmetries between the corpora exactly correlate with the language shift hierarchy in (5), but they still can be explained by sociolinguistic factors. Interestingly, our data do not confirm the previous observations on CS types and language shift.

1) *Inter-clausal switches*: Ulch > Nanai > Moksha > Hill Mari. The rate of clausal switches in the corpora correlates well with the language shift hierarchy: the more progressed language shift is, the more frequently clausal switches occur. At the same time, [Bentahila and Davies 1992] report the opposite tendency for code-switching between Moroccan Arabic and French. This contradiction can result from the deliberate specific of our texts. Being instructed to speak IL, speakers with dominant Russian (the Ulchas) try to speak IL, but insert Russian clauses in cases where they have difficulties with IL. Therefore, inter-clausal switches have to be rather frequent in their speech. However, in spontaneous communication, the same speakers would speak mostly Russian and include only short IL-fragments in their Russian speech, i.e. in fact use more intra-clausal switches. In contrast, balanced bilinguals (the Hill Mari speakers) do not need to use Russian sentences more often than sentences in IL, since they are equally prominent in both languages.

2) *Word-internal switches*: Moksha >> Nanai ≈ Ulch ≈ Hill Mari. The same apparent contradiction takes place for word-internal switches: their distribution corresponds neither to the hierarchy in (5), nor to the previous observations. Word-internal switches are mostly connected to cultural vocabulary (including “soviet realities”). According to [Bentahila and Davies 1992], they are frequent among the speakers with the dominant Arabic using French cultural words. In contrast, in our data, both speakers with dominant Russian (Ulch, Nanai) and balanced bilinguals (Hill Mari) use word-internal (Russian) switches with comparable frequency. In our corpora (in contrast to that of Bentahila and Davies), morphologically integrated cultural words come from the dominant language (Russian). For balanced bilinguals (Hill Mari) cultural words seem to be the main source of word-internal switches. In situations of progressed language shift (Nanai and Ulch), basic words are involved in CS as well as cultural lexicon, so the expected number of word-internal switches might be higher than it is. However, in the situation of progressed language shift, speakers are not very creative in IL-morphology and prefer to use non-integrated Russian constituents instead of morphologically-integrated Russian

stems, so the rate of word-internal shifts is as low as for balanced bilinguals. In contrast, on the intermediate stage of language shift the demand for Russian lexemes is equally high, but speakers feel free in integrating them into IL. This is the case of Moksha.

3) *Syntactic integration* (*constituent type*): Ulch > Nanai > Moksha > Hill Mari. For switched intra-clausal constituents, the degree of syntactic integration correlates with the language shift hierarchy: in the situation of language shift syntactically integrated constituents (e.g. NPs) tend to be switched, while balanced bilinguals more frequently switch non-integrated constituents (e.g. discourse markers). These two strategies of CS correlate with [Muysken's 2000] *insertion* and *alternation* respectively. According to Muysken, insertions are single constituents, content rather than functional words and complements rather than adjuncts. This is exactly what opposes NPs and PPs (counted as "integrated") to discourse particles, adverbs, interjections etc. (counted as "non-integrated"). Alternation, on the contrary, requires less integration into syntax, and is mostly represented by discourse particles, adverbs and other items, counted in our study as "non-integrated". Therefore, in Muysken's terms, languages more affected by language shift prefer insertion, while less affected ones prefer alternation.

Muysken himself [Muysken 2000: 227–228; 247–248] makes the general prediction, that in the process of language shift the rate of insertions would become higher and the rate of alternations lower (see also [Backus 1996] for the same claim). There is a nuance that has to be clarified. Making his prediction, [Muysken 2000] considers as insertions not only syntactically-integrated constituents, but also word-internal switches. However, we have already shown that word-internal switches are typical not for the progressed language shift situation (in contrast to syntactically integrated switches) and not for balanced bilinguals (as follows from Mysken's generalization), but for the early stage of language shift. The extensive use of word-internal and syntactically integrated switches is caused by the same reason, i.e. the need for Russian nouns. Ulch and Nanai speakers (progressed language shift) use switched NPs largely, since they are restricted in morphological adaptation of Russian nouns and syntactic integration is the only option in this case (see above). This leads to the high rate of syntactically integrated switched constituents. Moksha speakers (the early stage of language shift) mark Russian nouns with IL-affixes. This leads to the high rate of word-internal switches. In contrast, Hill Mari balanced bilinguals widely use IL nouns instead of Russian ones, so they show both the lowest rate of syntactically integrated and word-internal switches. Thus, in our data, the correlation between the degree of language shift and the rate of insertions, observed by Muysken, works differently, and an additional parameter of morphological vs. syntactic integration within the insertion should be considered.

Summing up, we can say that a very simple annotation, that contains only tags for languages and constituent types, can indeed shed light on correlation between sociolinguistic situation in the community and CS types. The metrics provide numeric data which can be projected on a hierarchy of language shift. Strong oppositions between situations with language shift and without it hold, however, they can work in the opposite direction as well. The explanation of such a variation is an object for a future research. A possible parameter that has to be considered is (non)-equivalence of the main language of the clause/text and the dominant language.

6. Abbreviations

3—3rd person, ACC—accusative, F—feminine, GEN—genitive, IN—inessive, IPFV—imperfective, N—neuter, NEG—negative, NPST—non-past, PL—plural, PRS—present, PST—past, PTCL—particle, R—Russian, SG—singular.

References

1. *Backus A.* (1996), Two in One: Bilingual Speech of Turkish Immigrants in the Netherlands, Tilburg University Press, Tilburg.
2. *Barnett R., Codo E., Eppler E., Forcadell M., Gardner-Chloros P., van Hout R., Moyer M., Torras M. C., Turell M. T., Sebba M., Starren M., Wensing S.* (2000), The LIDES Coding Manual: A document for preparing and analyzing language interaction data, Version 1.1–July, 1999, International Journal of Bilingualism, 4(2), pp. 131–132.
3. *Bentahila A., Davies E. D.* (1992), Code-switching and language dominance, Harris, R. J. (Ed.) Cognitive processing in bilinguals, Elsevier, Amsterdam, pp. 443–458.
4. *Gambäck B., Das A.* (2014), On measuring the complexity of code-mixing, Proceedings of the 11th International Conference on Natural Language Processing, Goa, India, pp. 1–7.
5. *Gambäck B., Das A.* (2016), Comparing the level of code-switching in corpora, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 1850–1855.
6. *Gardner-Chloros P, Moyer M., Sebba M.* (2007), Coding and Analysing Multilingual Data: The LIDES Project,Creating and Digitizing Language Corpora, pp. 91–120.
7. *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, pp. 246–257.
8. *Guzmán G. A., Serigos J., Bullock B., Toribio A. J.* (2016), Simple tools for exploring variations in code-switching for linguists, EMNLP-2016: Proceedings of the Second Workshop on Computational Approaches to Code Switching, pp. 12–20.
9. *Guzmán G. A., Serigos J., Bullock B., Toribio A. J.* (2017a), Moving code-switching research toward more empirically grounded methods, Proceedings of the Workshop on Corpora in the Digital Humanities (CDH 2017), Bloomington, IN, USA, pp. 1–9.
10. *Guzmán G. A., Serigos J., Bullock B., Toribio A. J.* (2017b), Metrics for modeling code-switching across corpora, Proc. Interspeech 2017, pp. 67–71.
11. *Goh K. I., Barabási A. L.* (2008), Burstiness and memory in complex systems, EPL (Europhysics Letters), 81(4): 48002.
12. *Kalinina E. Ju., Oskolskaya S. A.* (2016), Nanai [Nanajskij jazyk], Language and Society. An Encyclopaedia [Jazyk i obščestvo. Ènciklopedija], Azbukovnik, Moscow, pp. 293–296.

13. *Koryakov Ju. B., Kholodilova M. A.* (2018), General information about the Moksha language and the idiom [Obsčie svedenija o mokshanskom jazyke i issleduemom govore], Toldova S. Ju., Kholodilova M. A. (Eds.), Elementy mokshanskogo jazyka v tipologicheskom osveschenii, Buki Vedi, Moscow, pp. 6–18.
14. *Muysken P.* (2000), Bilingual speech: A typology of code-mixing, Cambridge University Press, Cambridge/New York.
15. *Myers-Scotton C.* (1992), Comparing codeswitching and borrowing, Journal of Multilingual & Multicultural Development, 13(1–2), pp. 19–39.
16. *Myers-Scotton C.* (1993), Duelling languages: Grammatical structure in codeswitching, Oxford University Press, Oxford/New York.
17. *Myers-Scotton C.* (2002), Contact linguistics: Bilingual encounters and grammatical outcomes, Oxford University Press, Oxford/New York.
18. *Sumbatova N. R., Gusev V. Ju.* (2016), Ulch [Ul'čskij jazyk], Language and Society. An Encyclopaedia [Jazyk i obščestvo. Ènciklopedija], Azbukovnik, Moscow, pp. 513–515.
19. *Treffers-Daller, J.* (1991), Towards a uniform approach to codeswitching and borrowing. Papers for the workshop on constraints, conditions and models, European Science Foundation, Strasbourg, pp. 259–279.