Annotating and exploring code-switching in four corpora of minority languages of 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 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 lunguistics, quantitative linguistic studies, language shift, code switching, code-switching metrics, Uralic languages, Tungusic languages

1. Introduction

The aim of this paper is to show the correlation between socio-linguistic factors and frequency and types of code-switching (CS) based on statistical metrics that can be calculated using corpora of oral discourse of bilingual speakers of several minority languages of Russia.

Four languages of Russia were chosen for the research: Moksha, Hill Mari (Uralic); Nanai, and Ulch (Tungusic). Almost all speakers are bilingual, but sociolinguistic situations in the communities are quite different. While across speakers of Hill Mari a stable balanced bilingualism takes place, the Moksha speakers, the Nanais and especially the Ulchas, are undergoing a progressing language shift to the dominant Russian language (Kalinina & Oskolskaya 2016; Gerasimova 2002; Sumbatova & Gusev 2016). This can be represented by a following hierarchy of language shift:

(1) Ulch > Nanai > Moksha >> Hill Mari

The use of two or more languages within one conversation or even within one utterance is known as the phenomenon of CS. In this project, we understand it quite broadly: from intersentential switching involving entire clauses (2) and intrasentential switching of constituents of different length (see *ixnjuju familiju* in (3)) to nonce borrowings from Russian (see sestra-ni [sister.R-3SG] in (3)), that often bear morphological affixes of the “main” language of the text.

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

(3)	*i*	*ti*	*sestra-ni*	*ti*	*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)

The corpora 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 IL. For all the narrators, the IL is the first one, or it was acquired simultaneously with Russian. All the speakers are highly proficient in Russian. They are also proficient in the IL enough to tell a spontaneous story. The crucial difference among them the current use of the languages. The Hill Mari speakers use the IL in their everyday communication at least as frequently as Russian. The Ulch narrators use the IL much more restrictively than Russian or do not use it at all. In the Moksha and Nanai samples the situation varies across speakers.

All the corpora contain transcriptions and Russian translations. We added a specific annotation of two types to them. The annotation is discussed in 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). We compared the information for our four corpora. The aim was to reveal specific properties of CS that vary across the text collections under discussion. The general hypothesis was that the main difference 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. The results of the study are discussed in 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, CS_TYPE, which indicates the syntactic type of the switched fragment (all tags are aligned to words).

The LANG tier contains two tags: IL – indigenous language and Rus – Russian (including Russian words with IL morphology). This tier was created semi-automatically: the tag Rus was assigned to all words transcribed in Cyrillic, but some tags were changed or added manually, since not all Russian fragments were written in Cyrillic in the original transcription.

The CS_TYPE tier contains tags listed in Table 1. These tags were assigned manually to a smaller part of the text collection (see Table 2 for the meta-information about corpora).

Table 1. The list of CS-type tags

tag	**description**
adj(+)[1]	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[2]	clause with Russian verb
voc(+)	vocative forms
vp(+)	verb phrase
other	other constituent types

Table 2. The corpora: size 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

Russian fragments that do not form any syntactic constituent are marked with corresponding tags separately (*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)

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

[2] 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.

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

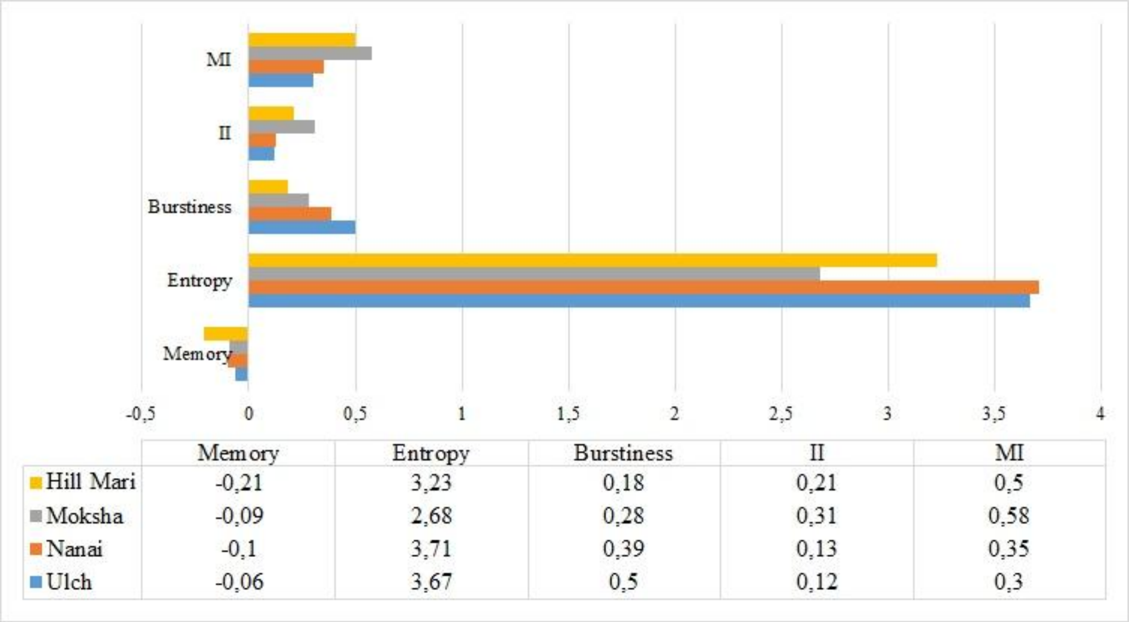
The general information on CS, which allows us to compare the four text collections, was obtained from the word-by-word language annotation, available for the whole corpora (cf. Section 2). To characterize CS patterns, we used the existing metrics, proposed for corpus-based studies on CS and summarized in 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 3.

Table 3. 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	$MI = \frac{1 - \sum p_j^2}{\sum p_j^2}$	[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	$II = \frac{1}{n-1} \sum_{1 \le i < j \le n} S(l_i, l_j)$	[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	$Span\ Entropy = -\sum_{l=1}^{M} p_l log_2(p_l)$	[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	$Memory = \\ = \frac{1}{n_r - 1} \sum_{i=1}^{n_r - 1} \frac{(\tau_i - m_1)(\tau_{i+1} - m_2)}{\sigma_1 \sigma_2}$	[-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.

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



	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

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 has the lowest number Russian words, while the Moksha corpus has the highest number (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.

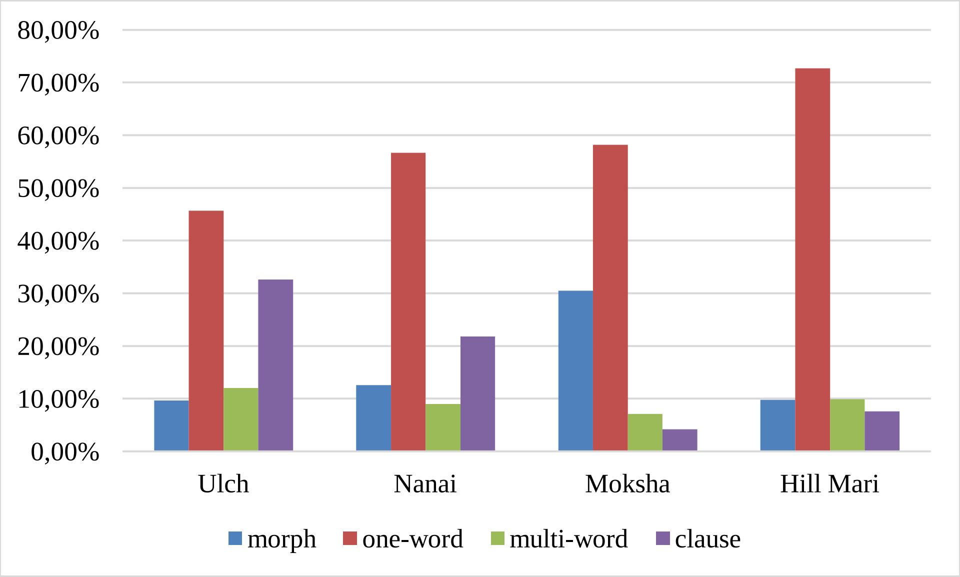
We also calculated metrics that include information about the temporal distribution of CS across the corpus using language spans – the distance between switch points, i.e. the length of monolingual discourse. The **Burstiness** measures whether CS has a periodic character or occurs in bursts. All our corpora have unpredictable patterns of switching with the following hierarchy: Ulch > Nanai > Moksha > Hill Mari. Ulch and Nanai contain less Russian fragments and are more unpredictable. The switching in Hill Mari is the most predictable, although Moksha has more Russian fragments than Hill Mari. 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. 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 influence the length of following spans. The hierarchy is as follows: Ulch > Nanai > Moksha > Hill Mari. All language spans are rather unpredictable, but Hill Mari language spans are more negatively correlated, while Ulch language spans are more positively correlated.

Thus, in our collection, the Finno-Ugric languages are opposed to Tungusic ones. The former have a better socio-linguistic situation, yet they show a higher number, a higher frequency and a higher predictability of CS.

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[3]) is given in Figure 2.

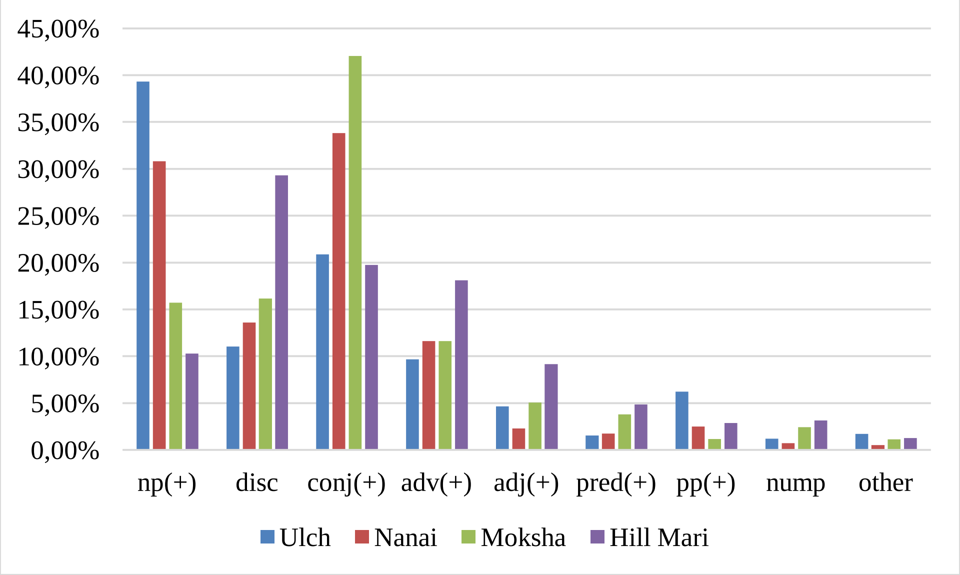
Fig. 2. Switched fragments' size



The most frequent type of CS for all the collections are one-word switches (from 45,66 to 72,67%). Multi-word constituents are, on the contrary, relatively rare. In contrast, the rate of word-internal switches and that of switched clauses differ a lot. The rate of word-internal switches is much higher for Moksha than for all other languages. The percentage of clausal switches in Ulch and Nanai is at least thrice as much in Moksha and Hill Mari.

For switched constituents (excluding Russian stems with IL-morphology and Russian sentences), we calculated the frequency distribution of different syntactic types, and the results are presented in Figure 3. Only frequent types are included (> 1%). The rest includes such types as *dep*, *ideoph*, *interj* and *vocp(+)*.

Fig. 3. Syntactic types of intra-sentential CS

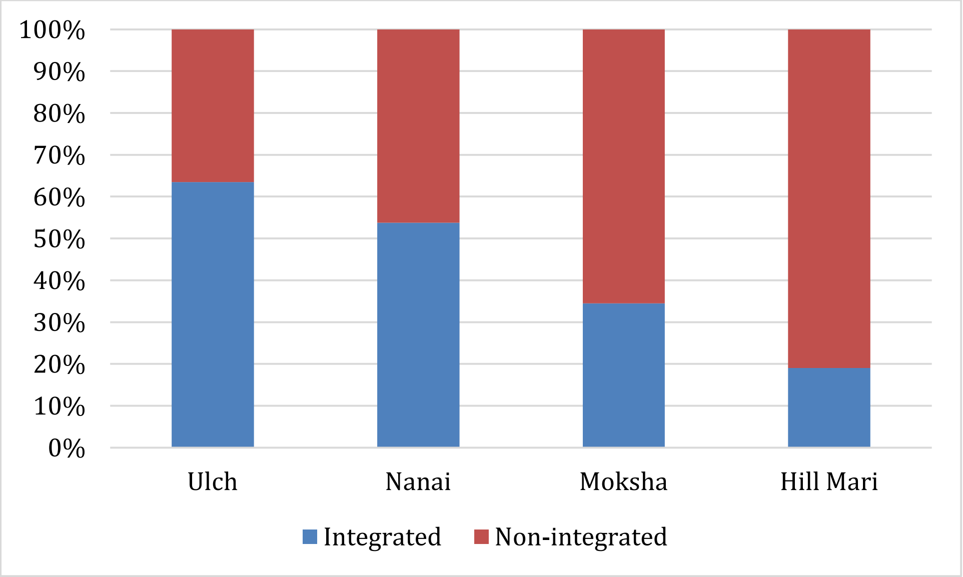


[3] Multi-clause switched fragments were not treated separately. Each of them was counted as several independent switched clauses. The same is true for multi-word switched fragments that do not form a syntactic constituent: they were counted as several independent constituents.

The most frequent types of switched constituents in all the corpora are *np* (nominal phrase), *disc* (discourse markers), *conj* (conjunction phrase, mostly sole conjunctions) and *adj* (adjectives) while other types are presented more or less regularly in a given language.

We expect the ratios to reflect the language shift hierarchy introduced in (1). Adverbs and 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 summary ratio for non-integrated and integrated elements where we treated constituents of types *np(+), pp(+), nump(+)* as integrated and *disc, pred(+), interj, adv(+)* and *adj(+)* as non-integrated (see Figure 4)[4].

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



The Figure 4 demonstrates that there is indeed a correlation between the degree of language shift and the number of integrated elements: the speakers with language shift (Ulch and Nanai) more readily insert integrated elements into their discourse.

5. Results and discussion

The crucial difference between the text collections under discussion concerns language dominance and the degree of language shift attested in the community.

The prediction was that the differences in CS would follow the same hierarchy of languages agreeing with this sociolinguistic difference.

We have checked the following macro-parameters of CS: Multilingual Index, Integration Index, Burstiness, Entropy and Memory. These measures helped us to check whether quantitative results are interpretable in terms of language shift hierarchy.

We also checked the following micro-parameters of CS: the rate of clausal switches, the rate of word-internal switches, the rate of different syntactic types of switched constituents, the rate of non-constituent clause-internal switches. Basing mostly on generalizations made in Benthalia & Davies (1992), Backus (1996), and Muysken (2000: 227–228; 247–248), we have the following expectations on correlations between language dominance and inter-generation shift[5], on the one hand and CS types, on the other hand:

[4] *conj*, which do not form part either of the two types of elements, were excluded.

[5] 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.

- 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) decreases, while the number of less integrated constituents (alternations in terms of Muysken 2000) increases;
- in the situation of language shift the number of non-constituent switches (which also belong to alternations in terms of Muysken 2000) increases.

The data show that the correlation between socio-linguistic situation and the most frequent types of CS can be not so straightforward, if not the opposite.

1) *Inter-clausal switches*. In the corpora, the highest rate of clausal switches is attested in Ulch, which is the most affected by language shift. Bentahila and Davies (1992), report the opposite tendency for Moroccan Arabic and French. This contradiction can be due to the deliberate specific of our texts. Nevertheless, even in such artificial conditions the percentage of switched clauses is considerably higher in the case of Nanai and Ulch than in the case of Moksha and Hill Mari. Thus, the languages from the opposite ends of hierarchy still differ crucially in this aspect.

2) *Word-internal switches*. The same apparent contradiction takes place for word-internal switches. They are mostly connected to cultural vocabulary (including “soviet realities”). In Bentahila and Davies (1992) they are frequent among the speakers with the dominant Arabic, and morphologically integrated cultural words come from French. In our corpora, morphologically integrated cultural words come from the dominant Russian, which however is not the ML in this type of texts. Thus, the attested distribution seems natural. Some Russian cultural words are needed both by balanced bilinguals and speakers with dominant Russian. At the same time, speakers tend to keep IL word forms as frozen items. This explains quite a modest number of word-internal switches in Nanai and especially in Ulch. In contrast, on the intermediate stage of language shift the demand for Russian lexemes is equally high, but speakers feel free in their marking in IL affixes. This is the case of Moksha. Speakers of Moksha and Hill Mari often do not use Russian inflectional morphology which accounts for a large number of inserted adjectives – even if cultural words are used the agreement between noun and adjective is usually absent (*čeboksarsk-**ij** ges* (masc.) instead of *čeboksarsk-**aja** ges* (fem.) ‘Cheboksary hydroelectric power plant’).

3) *Constituent type*. We showed that the constituents can be divided into two groups: integrated and non-integrated. The two tendencies observed in Tungusic and Finno-Ugric can be interpreted as two strategies of CS, that 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 to conjunctions, discourse particles, adverbs and adjectives. *Alternation*, on the contrary, requires less integration into syntax, and is mostly represented by discourse particles, adverbs and adjectives. Therefore, Ulch prefers insertion to alternation, while Hill Mari and Moksha seem to prefer alternation which can be seen by the high percentage of adjectives, adverbs and discourse markers. Again, the correlation is the opposite.

There is another subtle nuance that has to be considered. The Table 6 shows that the most frequent type of switched elements in Moksha are roots (*morph*). As pointed out by Muysken (2000), morphological integration can also be an indicator of insertion. However, Muysken makes the opposite prediction, that in the process of language shift the rate of insertions would become higher and the rate of alternations would become lower (see also (Backus 1996) for the same claim). According to our data, this is not exactly so, since word-internal switches (which are insertions in Muysken’s terms) are very frequent in case of Moksha but not so frequent in case of Nanai and Ulch. However, concerning Tungusic languages, the tendency revealed in our data seems logical. Ulch speakers with highly dominant Russian use more syntactically integrated constituents (insertions), since they need to use Russian nouns in argument position and they are restricted in their morphological adaptation, so syntactic integration is the only option. Moksha speakers mark nouns with IL affixes, and Hill Mari bilinguals widely use IL nouns. The question of what determines the facility of insertion in case of Moksha needs however further investigation. Nevertheless, the tendency noticed by Muysken does not seem universal: different types of items attributed by him as insertions vs. alternations behave differently.

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 socio-linguistic 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 ML and the dominant language in the text collection.

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

Backus A. (1996), Two in One: Bilingual Speech of Turkish Immigrants in the Netherlands, Tilburg University Press, Tilburg.

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.

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.

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.

Penelope Gardner-Chloros, Melissa Moyer, Mark Sebba. 2007. Coding and Analysing Multilingual Data: The LIDES Project // Creating and Digitizing Language Corpora pp 91-120.

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.

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.

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.

Guzmán G.A., Serigos J., Bullock B., Toribio A. J. (2017b), Metrics for modeling code-switching across corpora, Proc. Interspeech 2017, pp. 67–71.

Goh K.I., Barabási A. L. (2008), Burstiness and memory in complex systems, EPL (Europhysics Letters), 81(4): 48002.

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

Muysken P. (2000), Bilingual speech: A typology of code-mixing, Cambridge University Press, Cambridge/New York.

Myers-Scotton C. (1993), Duelling languages: Grammatical structure in codeswitching, Oxford University Press, Oxford/New York.

Myers-Scotton C. (2002), Contact linguistics: Bilingual encounters and grammatical outcomes, Oxford University Press, Oxford/New York.

Sumbatova N.R., Gusev V.Yu. (2016), Ulch [Ul’čskij jazyk], Language and Society. An Encyclopaedia [Jazyk i obščestvo. Ènciklopedija], Azbukovnik, Moscow, pp. 513–515.